2011

Introduction to GIS Using Open Source Software, 1st ed

Frank Donnelly
CUNY Bernard M Baruch College

Introduction to GIS using Open Source Software

Frank Donnelly, Geospatial Data Librarian, Baruch College CUNY

francis.donnelly@baruch.cuny.edu

Last Updated: January 26, 2011 (Version 1.0)

This tutorial was created to accompany the GIS Practicum, a day-long workshop offered by the Newman Library at Baruch College CUNY that introduces participants to geographic information systems (GIS) using the open source software QGIS. The practicum introduces GIS as a concept for envisioning information and as a tool for conducting geographic analyses and creating maps. Participants learn how to navigate a GIS interface, how to prepare layers and conduct a basic geographic analysis, and how to create thematic maps.

This tutorial was written using QGIS version 1.5 "Tethys", a cross-platform (Windows, Mac, Linux) desktop GIS software package. Shortly after it was written QGIS 1.6 "Copiapo" was released, and it contains a few updates over the previous version. Differences between versions are noted in the text whenever relevant. You can download the software and user manual from the QGIS website at http://www.qgis.org/. The data used for the tutorial can be download from the GIS Practicum page under the Tutorials and Courses tab on the Baruch GIS subject guide at http://guides.newman.baruch.cuny.edu/gis. Once you download and unzip the file, you'll see that the data files are separated into different folders for each part of the tutorial.

This tutorial and associated screenshots were created using QGIS in a Windows operating system. The names of certain tools and menus may vary slightly between operating systems, but functionality should be the same.

This document is divided into five parts and several subsections. Each subsection begins with steps for learning a specific application or process (the what and when), followed by commentary that explains various facets of the process (the how and why). The process and the commentary were separated in order to keep the steps as concise and easy to follow as possible with few digressions; you follow the steps first, and then go back and understand the details of why you followed the steps you did.

Objectives

Participants will be able to bring both the tools and the knowledge they gain from this workshop to enhance their projects and the organizations they work for. Specifically, this workshop will enable participants to:

- Add data to GIS software and navigate a GIS interface
- Perform basic geoprocessing operations for preparing vector GIS data
- Convert text-based data to a GIS data format
- Conduct geographic analyses using standard GIS tools and vector data
- Create thematic maps using the principles of map projections, data classification, symbolization, and cartographic design
- Locate GIS data on the web and consider the merits of different data sources
- Demonstrate competency with a specific GIS package (open source QGIS)
- Identify other GIS topics (tools and techniques for analysis), data formats (raster, vector), and software (open source and ArcGIS) to pursue for future study

Outline

- Part 1: General introduction and overview of GIS
- Part 2: Introduction to GIS Interface (learn how to navigate the interface: adding data, layering data, symbolization, changing zoom, viewing attributes, viewing attribute table, making basic selections, difference between data formats, organizing projects and data)
- Part 3: GIS Analysis (using site selection example in NYC, basic geoprocessing tasks, attribute table joins, plotting coordinate data, buffers, basic statistics, advanced selection)
- Part 4: Thematic mapping (using US states as an example, map projections, coordinate systems, data classification, symbolization, calculated fields, labeling, map layouts)
- Part 5: Going Further with GIS (exploring and evaluating online sources for free data, exploring open source and ArcGIS software resources for learning more)

Table of Contents

3. Creating Calculated Fields
    - Steps
    - Representing Values
    - Industrial Classification: NAICS
4. Classifying and Symbolizing Data
    - Steps
    - Data Classification and Color Schemes
    - New Symbology Tab
5. Designing Maps
    - Steps
    - QGIS Map Composer: Some Details
    - General Map Design
    - Output Formats
6. Adding Labels
    - Steps
    - Labeling in QGIS
    - Thematic Maps and Symbols
    - Considerations and Next Steps

5. Going Further
    1. Finding Data
    2. Data Sources
    3. Additional Concepts and Applications

Part 1 - An Overview of GIS

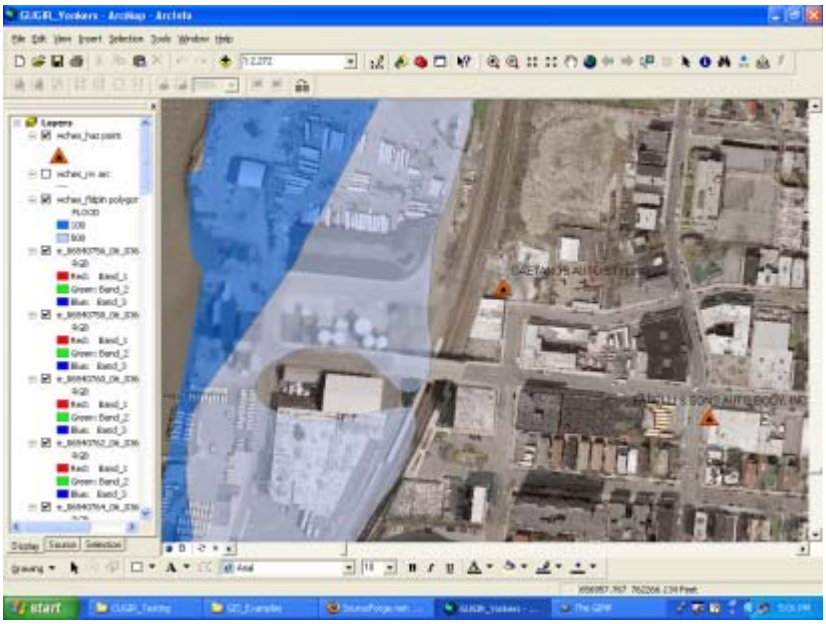
The goal of part 1 is to provide you with a basic foundation in GIS concepts and software in preparation for the rest of the tutorial.

Section I: Basic GIS Concepts

Geographic Information Systems (GIS) are an integrated collection of software and data used to visualize and organize geographic data, conduct spatial analysis, and create maps and other geospatial information. Narrow definitions of GIS focus on the software and data, while broader definitions include hardware (where the data and software is stored), metadata (data that describes the data), and the people who are part of the system and interact with it as creators, curators, and users.

Another definition: GIS is a visual system that organizes information around the concepts of place and location that can be used for geographic analysis, map making, database management, and geospatial statistics. GIS can be (and has been) applied to virtually any discipline or endeavor.

In a GIS, geographic features are represented as individual files or layers that can be added to a map. These features are not maps in and of themselves, but are the raw materials used for map making and analysis. For much of the 20th century cartographers drew geographic features on individual mylar or acetate sheets and then layered those sheets over a paper base map to create maps. GIS uses the same principles of layering, with individual files consisting of features that can be layered on top of each other in GIS software. GIS software acts as an interface, or window, for viewing and manipulating GIS data. The ability to add different layers is quite powerful, as combining the layers allows for analysis that would be impossible if you were viewing single layers by themselves (see example below of air photo, flood zones, and hazardous site layered):

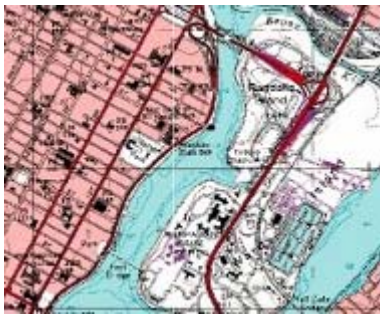
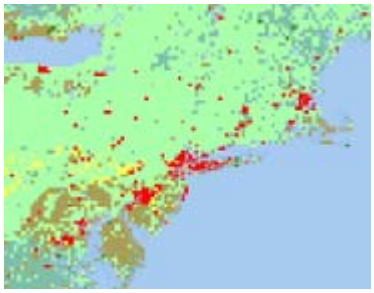


Each GIS file is georeferenced, meaning that the file is actually tied and related to real locations on the earth. Just as paper maps were drawn based on map projections and coordinate systems, each GIS file has also been created based on a particular projection and coordinate system, which means that files that share the same reference systems can be laid on top of each other. Since projections and coordinate systems are highly

standardized, GIS data can easily be shared. If two files do not share the same system, most GIS software can convert files from one system to another so they'll match. This distinguishes map making in GIS versus a graphic design package. Maps created in a graphic design package are just simple lines and shapes with no connection to the earth, and the components of the map can't be easily replicated to make other maps. GIS files used to create maps in a GIS package can readily be shared and used to create any map, because they are tied to the earth using standardized systems.

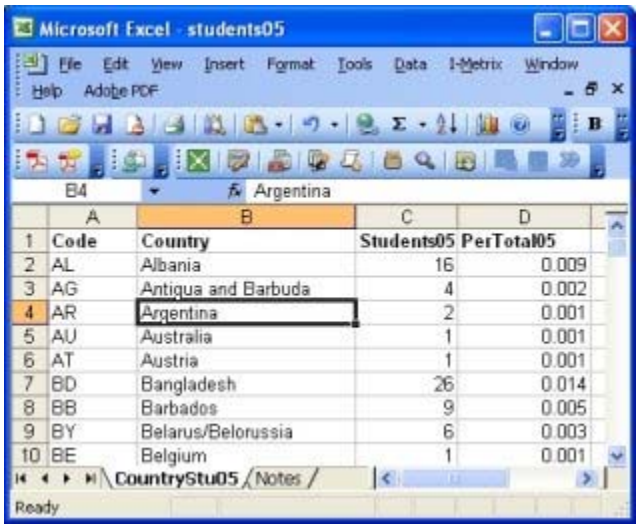
GIS files are stored in several formats, and each format comes in several different file types. Major formats and file include:

- Raster - represent a continuous surface that is divided into grid cells of equal size. Each cell appears as a particular color based on some value (i.e. reflected light). Files in the raster format are similar to digital photos. Common raster objects include air photos, satellite imagery, and paper maps that have been scanned. Raster files can also consist of photos or imagery that have been generalized or have had value added to them to create a new layer, like a land use and land cover layer or a grid showing temperature. There are many different file formats, some common ones include Tiffs (.tif), JPEGs (.jpg), and SID (.sid). Unlike regular .tif or .jpg files, GIS raster files are georeferenced.

- Vector - consists of discrete coordinates and surfaces that are represented as individual points, lines, or polygons (areas). Vector files appear to be more "map-like", and are always abstractions rather than actual images (i.e. shapes to represent boundaries, points to represent cities). Common file formats are ESRI shapefiles (.shp) ESRI coverages (.cov), Google KML files (.kml), and GRASS files.

- Tables - data tables that contain records for places can be converted to GIS files and mapped in several ways. If the data contains coordinates like latitude and longitude, the data can be plotted and converted to a vector file. If each data record contains unique ID codes for each place, those records can be joined to their corresponding features in a GIS file and mapped. Tables are commonly stored in text files like .txt or .csv, database files like .dbf, or in spreadsheets like Excel.



	A	B	C	D
1	Code	Country	Students05	PerTotal05
2	AL	Albania	16	0.009
3	AG	Antigua and Barbuda	4	0.002
4	AR	Argentina	2	0.001
5	AU	Australia	1	0.001
6	AT	Austria	1	0.001
7	BD	Bangladesh	26	0.014
8	BB	Barbados	9	0.005
9	BY	Belarus/Belorussia	6	0.003
10	BE	Belgium	1	0.001

CountryStu05 / Notes /
Ready

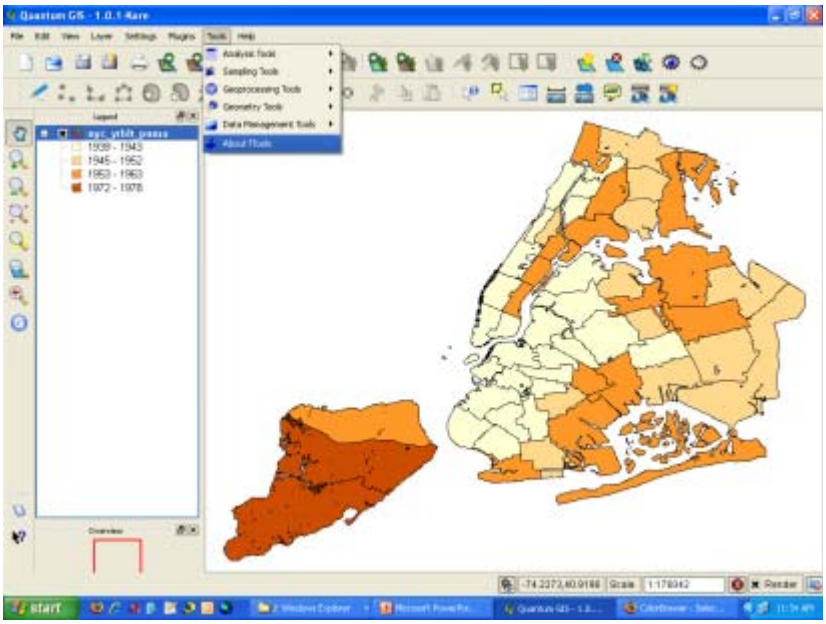
- Geodatabases - containers that can hold related raster, vector, and tabular data in one place. They are good for consolidating and organizing data. Geodatabases can be desktop (Microsoft Access .mdb, ESRI file geodatabases .gdb, Spatialite files .sqlite) or server based (PostGIS, ArcSDE).

Raster and vector GIS files exist spatialy, in that you can see the grid or shapes and their corresponding location on the earth, but also exist in tabular form. This is particularly valuable in the case of vector files. For example, every feature in a vector file showing country boundaries has an attribute table attached to it that has a record for each country. This attribute table contains columns or fields that store values for each country, such as the country's name, values like population or area that describe it, and ID codes that uniquely identify each one. The names can be used by the GIS to label each country, and the values like population can be thematically mapped.

The ID codes for each country can be used to join the attribute table for the GIS file to a tabular file that contains country-level data. For example, a GIS file of country boundaries with a country code can be joined within the GIS using relational database techniques to a text or spreadsheet file that has country-level data and that uses the same country codes to identify each country. The data in the table, which was just a regular table with no geospatial geometry, can now be visualized and mapped in GIS. There a number of standard ID codes that have been created which can be used for joining data. The two most common families of codes are FIPS (created by the US government to identify every single geographic entity in the US; there are also FIPS codes for countries) and ISO (created by the International Standards Organization to identify countries and their subdivisions).

Section II: GIS Software

A standard interface for GIS software has evolved over time. Typically, GIS software has a data view that consists of a table of contents that lists files that have been added to a project, a data window that displays the GIS files, and a set of toolbars and menus for accessing various tools and launching various processes. Dragging the layers in the table of contents changes the drawing order of the layers, and right or left clicking on a layer in the table of contents will reveal individual properties for that particular feature. You can also access the attribute table of the feature and a symbol tab for changing how the features are depicted or classified. There are several tools for zooming in and out to examine different layers and to change the extent of the view.



The way that coordinate systems and projections are handled is different for individual GIS software packages. In general, the options are: define the projection and coordinate system for the project before adding the files, or the project automatically takes the projection of the first file added. If you try to add GIS files that have different projections, some software may try to re-project the data on the fly, while others will simply fail to draw the new layers. Even if the software can correctly draw a layer without the user defining it, or even if it can re-project layers on the fly, users will run into problems later on when trying to manipulate the GIS files. You should always be sure to define the projection properly and make sure that all files share the same one - most GIS software will give you the ability to re-project data.

GIS software provides users with a variety of ways for querying geographic data, either by selecting records in the attribute table or shapes in the view, or by conducting searches where you build queries to high-light features that contain specific attributes, or that have some relationship with another geographic layer.

GIS software comes with a variety of editing tools that allow you to modify the geometry of GIS files. For example, you can merge features together, break them apart, or clip out or select certain areas to create new files. Collectively these processes are known as Geoprocessing. You geoprocess layers in order to prepare raw data for analysis, to create new layers or data, or to simplify layers for cartographic or aesthetic purposes. GIS also provides the ability to edit files on a feature by feature basis.

Most GIS programs have a separate map layout or print layout, where the user can create finished maps with standard map elements like titles, legends, scale bars, north arrows, and accompanying text. Finished maps can be exported out of the GIS as static files, such as pdfs or jpgs.

Users can always save their GIS projects in a GIS project file. The scale and extent of the data view, symbolization and classification assigned to layers, map layouts, and links to GIS files used in the project are stored in the file. It's important to understand that the GIS files themselves are NOT stored inside the project file - the GIS data and the GIS project file exist independently. When adding data to a GIS, you are establishing a link from the GIS project to the GIS data - the GIS data is not stored within the project. Furthermore, changing the colors of the features or classifying them in a certain way has NO EFFECT on the actual GIS data files themselves. When you change symbols, you are only changing how the GIS program views the data - you're not changing the data itself.

This is an important concept to grasp. Essentially, the GIS software acts as a window for viewing and working with GIS data, which is stored outside the window. The GIS project file essentially stores the window dressing, of scale and symbolization. You never actually change the GIS data unless you go into an edit mode or conduct an operation that creates a new GIS file. This relationship is of crucial importance when it comes time to move or share files - if you move your project file or your data, the links between them will become broken, and you'll

need to re-establish the location between the project and the data in order to repair your project file.

Section III: Open Source

In this tutorial we will be using QGIS, which is free open source software (FOSS). Open source software is an alternative to proprietary software:

- Open source software is free - you don't have to purchase it, and you can freely distribute it to anyone else, as opposed to proprietary software which you must purchase and is copyrighted so that you typically can not share it with anyone.
- The source code, or actual computer programming, that was used to create the software is transparent, as opposed to proprietary software where the code is hidden and encrypted.
- Under the open source model the programming code is transparent and you are free to change and make improvements to it; this is strictly prohibited with proprietary software.

Open source software can be created in several ways. A programmer or developer creates software from scratch, because they have some need that isn't being met by current software. Over time, as other programmers discover the project they may choose to contribute to building or improving this software, and they rally around the creator and begin to form a group that becomes devoted to the project. The Linux operating system and the Perl programming languages essentially began this way. Alternatively, a group of people who receive support from a business or entrepreneurs take software that was formerly proprietary but is no longer viable, and they build on this product and re-release it as open source. The Mozilla Firefox browser (formerly the proprietary Netscape) and Open Office (formerly the proprietary Star Office) are examples of the latter.

Why would people want to bother with creating FOSS software?

- It gives programmers a chance to practice their skills
- It gives programmers a way to enhance their prestige for their craft, as they can become known in different programming circles
- Open source is an ethos for some, who believe that software and information should be free
- Some see it as a superior model - since the code is open, there is a better chance that improvements can be made more quickly and that bugs can be discovered more easily than in proprietary software, as open source harnesses the power of the masses
- Businesses may prefer it because it does not tie them to costly, proprietary software they may go out of date or out of business - with open source there is always someone who can take over a project and keep it going

The number of FOSS GIS packages has grown over the course of the last decade. In this tutorial we will be using Quantum GIS (QGIS), which was initially developed by a group of volunteers in 2002 as a simple GIS viewer but has evolved into one of the premier FOSS GIS packages.

The advantage of using QGIS for this tutorial: it's free, you can download it yourself if you have your own computer, it runs on any operating system, it is mature enough that it supports most essential GIS tasks plus a few intermediate and advanced ones, and it's relatively easy to use.

The disadvantage is that QGIS can't do everything that proprietary software can, is still working out some bugs, and doesn't have the name recognition that software like ArcGIS or MapINFO do. There also isn't as much in the way of documentation or tutorials for QGIS relative to the other options, but this is changing.

Open software tends to be modular rather than monolithic; you often have several, independent software

applications to perform different functions, rather than one, large piece of software that does it all. A typical FOSS GIS workstation may include several applications like QGIS (for viewing data, basic analyses, map making, generally working with vector data), GRASS (a more advanced GIS for doing analyses and modeling and for working with raster data), GDAL / OGR (command line tools for converting files and projections and for basic queries), and a geodatabase application (PostGIS for server-based databases and Spatialite / SQLite for desktop use).

ArcGIS, created by a company called ESRI, has been on the market for several decades and is the dominant, proprietary (non-FOSS) GIS software on the market. It's used by most government agencies and universities. Since it is rather expensive to purchase for individual use, you tend to see it more often in institutional settings. If you are affiliated with a college or university, chances are you'll be able to access it somewhere on your campus. ESRI does distribute trial versions of the software for education and home use. A rival product, MapINFO created by Pitney Bowes, has a smaller but equally dedicated following. If you find that you need to learn one of these products, making the transition from FOSS is relatively straight forward as most GIS software operate under the same properties and principles and share similar user interfaces.

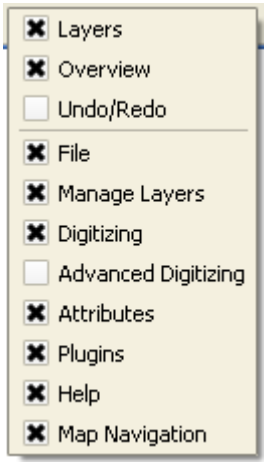
Part 2 - Exploring the Interface

The goal of part 2 is to familiarize you with the interface and basic features of GIS in general and QGIS in particular. You'll also add and configure some layers that you'll use later in Part 3.

Section I: The QGIS Interface

This section will introduce you to the QGIS interface; you will configure the interface in preparation for the rest of this tutorial.

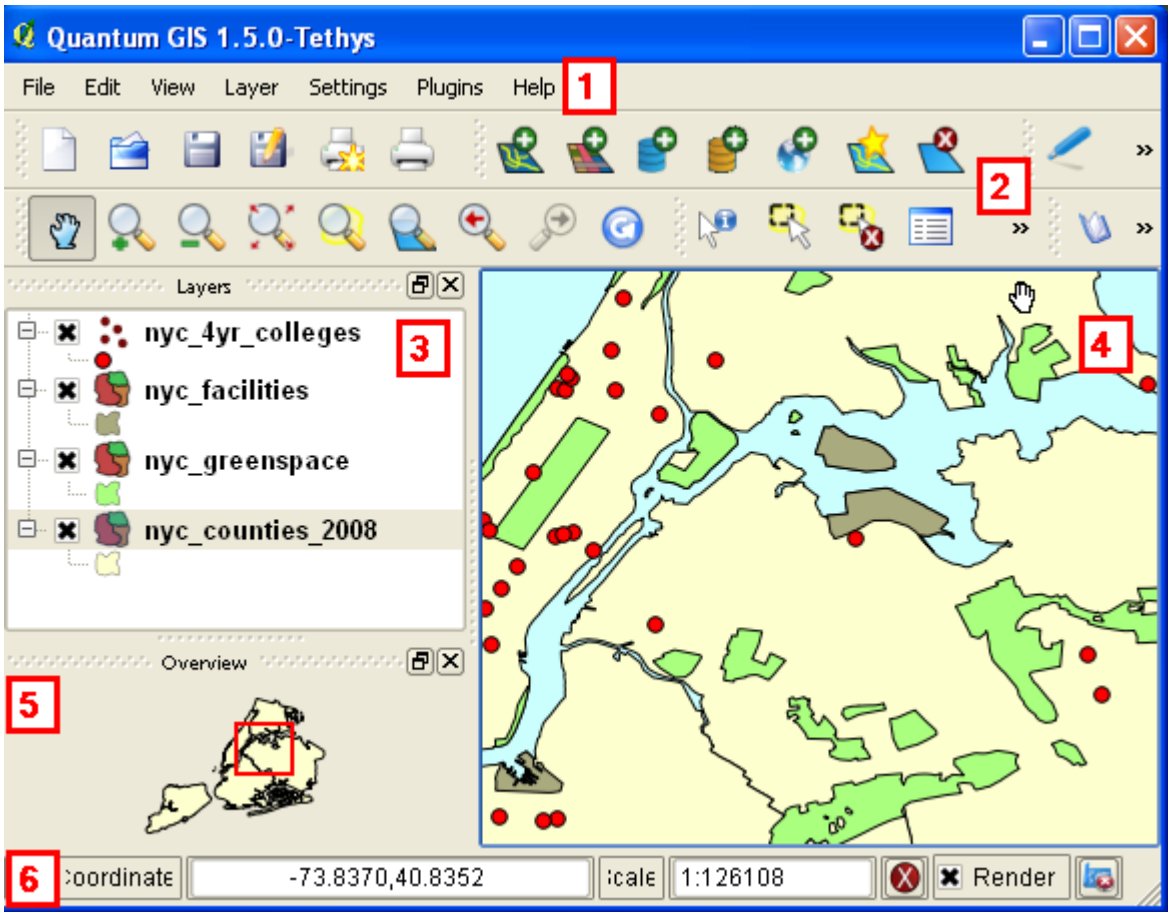
Steps



1. *Configure plugins.* Go to Plugins > Manage Plugins. Make sure the following three plugins ARE checked: Add Delimited Text layer, GdalTools, and ftools. Turn off all the other plugins.
2. *Configure the toolbars.* Right click on a blank area of the tool bar to get the tool bar view menu. Make sure the following two features are NOT checked: Undo / Redo and Advanced Digitizing. Make sure all of the other options are checked.
3. *Move toolbars.* Move the toolbars around by hovering over the left edge of a toolbar until you see a crosshairs, left click and hold, then drag and drop. Configure the toolbars to your liking (suggestion: try aligning them so you have only two rows of them at the top of the screen and all buttons are visible)

Commentary

1. *Menu Bar:* provides access to various features and functions of the software using a standard hierarchical menu. The location of the menus and menu items is fixed, although if you activate certain plugins they may add an additional menu to the bar.
2. *Tool Bar:* replicates many of the features and functions in the Menu Bar, providing access to common features in a single click. The location of the toolbars is not fixed; if you hover over the edge of the toolbar and hold down the left mouse button you can drag and dock the toolbar wherever you like (this means that the location of tools on your screen may not match those of other screens, or this tutorial).
3. *Map Legend:* a list of the map layers that are part of your current project. You can check or uncheck layers to turn them on and off, drag them to change the drawing order, select one in order to perform specific tasks on that layer, and right click on a layer to access menus and tools for working with that specific layer.
4. *Map View:* geographic display that shows all of your active layers.
5. *Map Overview:* you can add a layer to this overview to act as a frame of reference for the layers in your map view. It shows the full extent of a layer and outlines the portion of the area currently visible in the Map View in red.
6. *Status Bar:* shows the current scale of the map view and the coordinates of the current position of the cursor. Progress meters and other messages will appear here as you perform specific operations.



- *Want to turn a toolbar off? Wondering where a toolbar went?* If you right click on a blank area of either the Menu Bar or the Tool Bar, you'll get a list that shows all of the toolbars, as well as the Map Legend and Map Overview. You can check and uncheck items to turn them on and off.
- *Can't figure out what a button means or does?* If you hover over a button, a small window appears that displays the name of the button. If you select the What's This button and click on any area or item in the interface, you'll get a brief explanation of what it does.
- *Are there hotkeys?* Most menu items and tools can also be accessed by using hotkeys or keyboard shortcuts (for example, CTRL S will save the current project). For a full list of hotkeys, view the QGIS manual. Many of the common Windows shortcuts (like CTRL C for copy and CTRL V for paste) will work in QGIS.
- *Where is the QGIS manual?* These are available on the QGIS website at http://www.qgis.org/en/documentation/manuals.html.

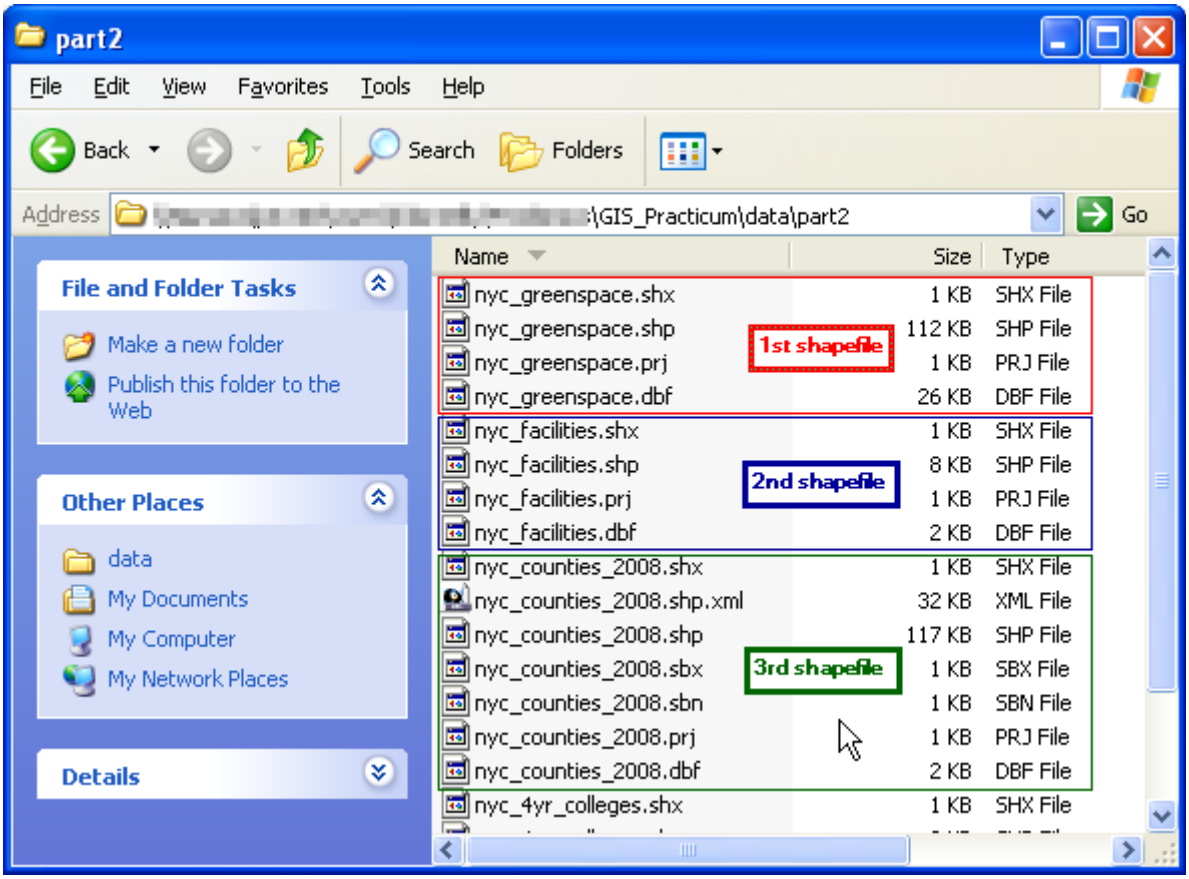
Section II: Adding Vector Data

In this section you'll learn how to add vector GIS files (shapefiles) to QGIS and to symbolize them. Shapefiles are a common GIS data format that you'll encounter in your future work.

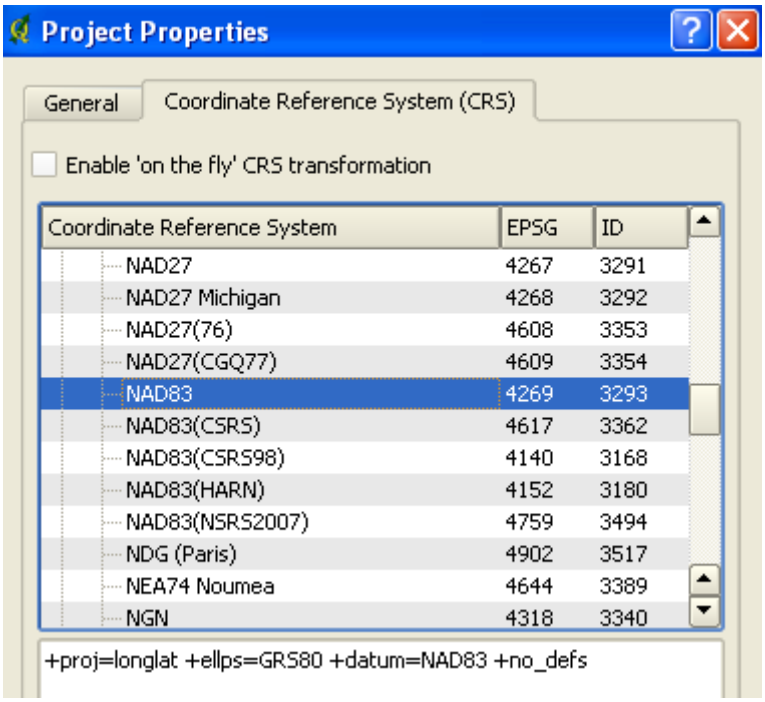
Steps

1. *Examine your data*. Take a look at the data files under the data folder for part 2. These are shapefiles that we will add to QGIS and work with for this project. There are four shapefiles; each shapefile is composed of multiple files

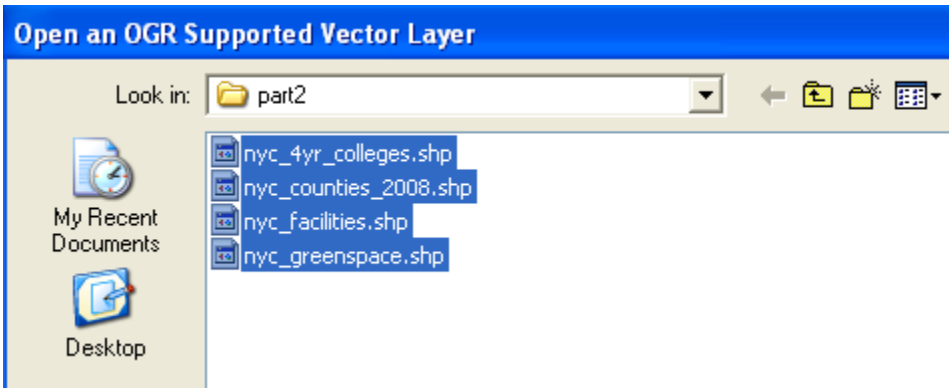
that have the same names but different extensions.



2. *Launch QGIS*. (If you're using Microsoft Windows, look under the Start Menu > Program Files > Quantum GIS > QGIS).
3. *Set the projection for the project.* On the Menu Bar, go to Settings > Project Properties > Coordinate Reference Systems Tab. Scroll through the list, choose NAD83, and hit OK (for now we'll just do this step and move on; we'll discuss coordinate systems and map projections later).



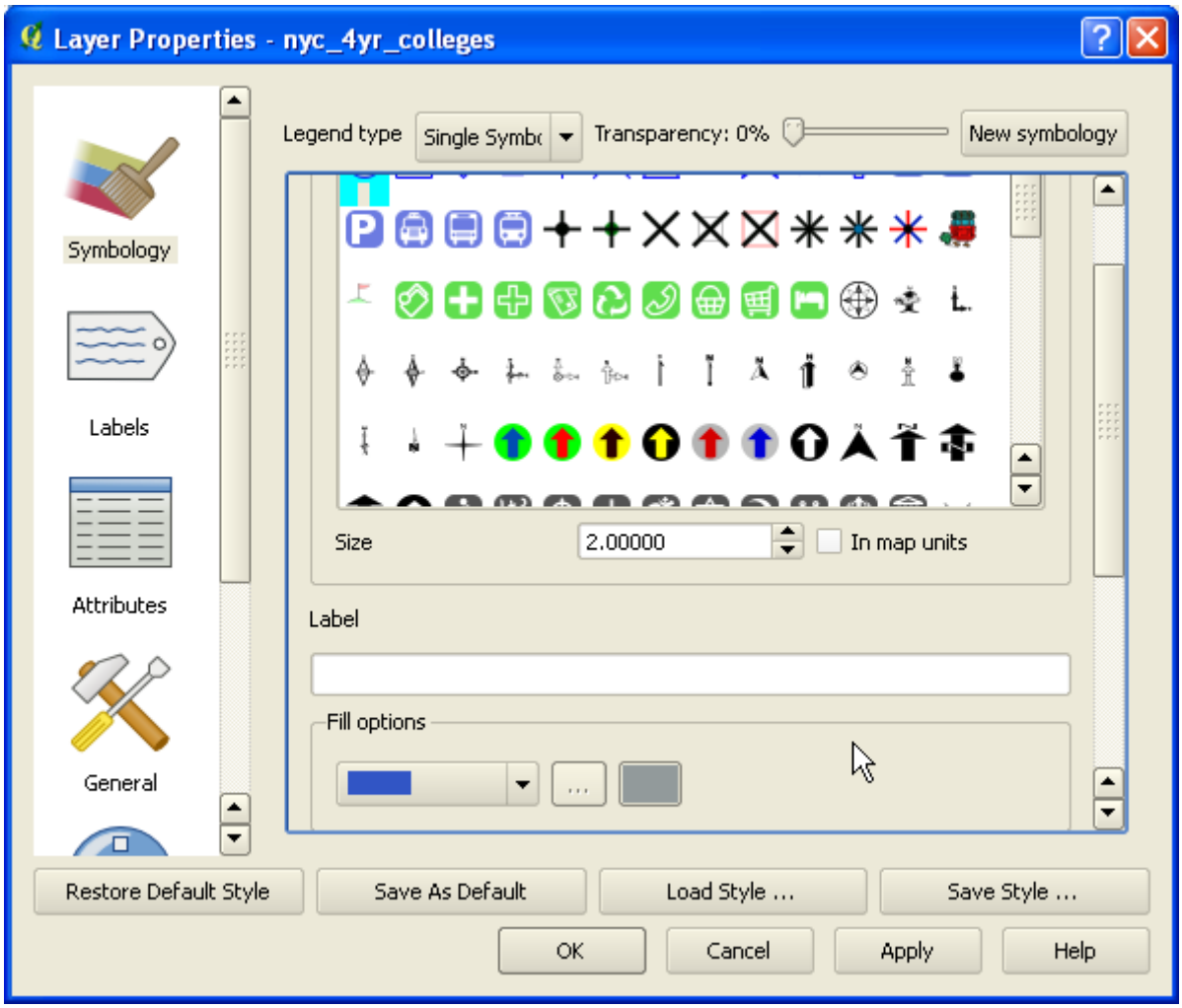
4. *Add the four shapefiles.* On the Tool Bar, hit the  Add Vector Layer button. When the Add Vector Layer box appears, hit the Browse button. Browse through the folder list to the data folder for part 2. Select the first layer in the list, hold down the shift key, then select the last layer. This should select all four shapefiles. Hit Open to add them. Your layers should appear in the Map Legend and Map View.



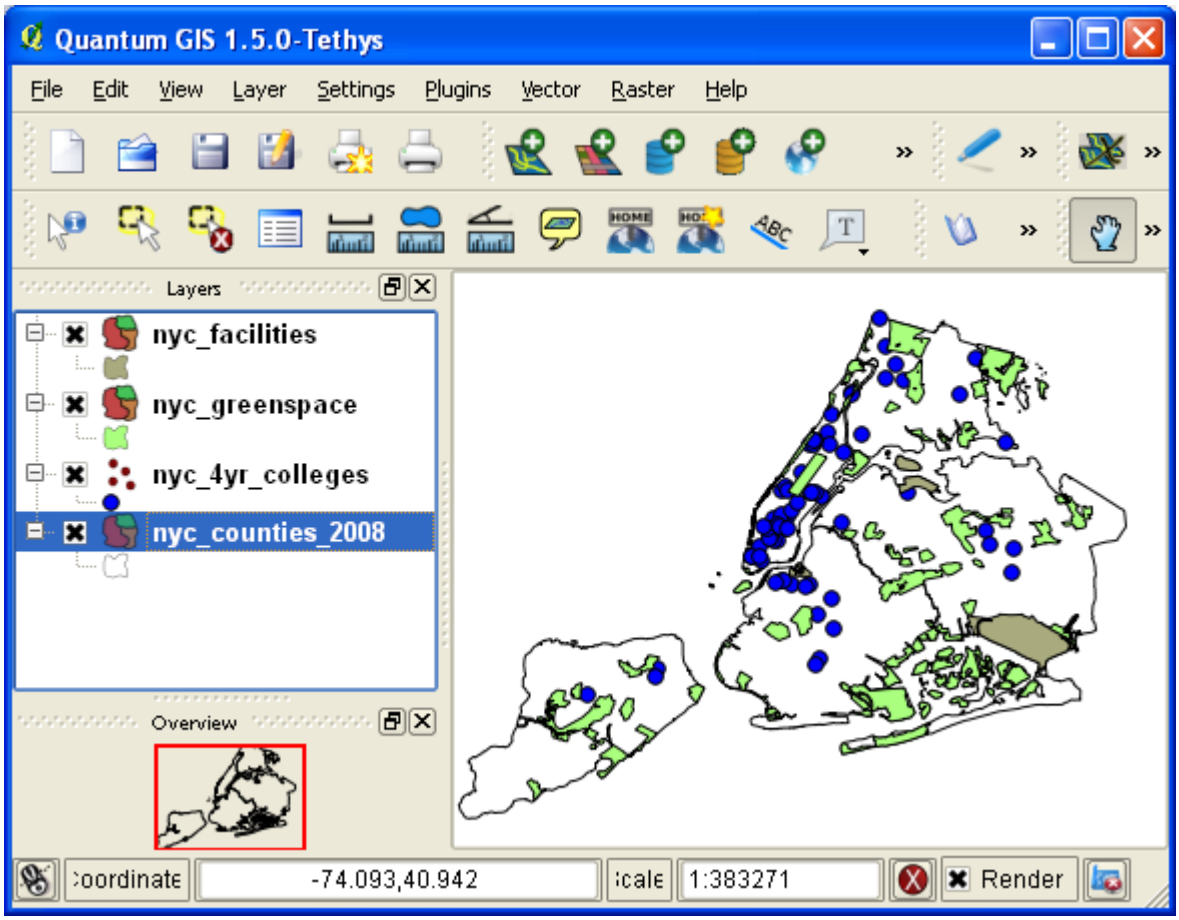
5. *Do your layers look jagged?* If not, skip this step. If so, on the Menu Bar, select Settings > Options > Rendering SVG, and under Rendering Quality check the box that says "Make lines appear less jagged at the expense of some drawing performance", and hit OK.

6. *Experiment with changing the drawing order.* Click on the first layer that's listed in the Map Legend (ML), hold down the left mouse button, and drag it to the bottom of the list. This moves that layer from the top of the drawing order to the bottom; layers in the Map Legend (ML) are stacked on top of each other, and their order in the list determines which are visible relative to others. Move the counties layer to the top of the list to see what happens.

7. *Order the layers.* Drag the layers in the Map Legend (ML) so they appear in this order, from top to bottom: nyc_4yr_colleges (colleges and universities), nyc_greenspace (parks and wildlife areas), nyc_facilities (airports, ports, prisons), nyc_counties_2008 (counties) / boroughs).
8. *Change the color for the colleges.* Double-click on the colleges layer in the ML to open the Layer Properties menu for that layer. Click on the Symbology tab. Click on the box under Fill options that contains the fill color. Change the color to blue by choosing a box in the color pallete. Click OK, then OK again on the Symbology menu.



9. *Change the colors for parks and facilities layers.* Make the parks green and the facilities grey or brown.
10. *Give the counties no fill.* (i.e. make them hollow with no color). Double-click on the counties layer in the ML to open the Layer Properties menu for that layer. Click on the Symbology tab. Change the option in the Fill Options drop down box to None. Click OK.
11. *Add the counties layer to the overview.* Select the counties layer in the ML, right click on it and click on the Add to Overview option in the menu. After completing these steps, your QGIS window should resemble the image below.



Commentary

Shapefiles

A shapefile is a very common file format used for storing vector GIS data. It was created by a company called ESRI, the makers of ArcGIS (the predominant software in the proprietary GIS market). Shapefiles are an open GIS format that can be used in just about any GIS software package, including QGIS. A shapefile can consist of point, line, or polygon features for a given geographic area, and can never consist of multiple types of geometry (i.e. you can't have a shapefile with points and lines).

Despite it's singular sounding name, a shapefile consists of several individual files. The following three pieces are mandatory:

- .shp file - shape file, contains the geometry
- .shx file - shape index file, an index of the geometry
- .dbf file - attribute file, contains attributes for the features

The following pieces are typically (ideally) included

- .prj file - a plain text file that contains the projection and coordinate system
- .sbn and .sbx files - spatial index of the features
- .shp.xml file - XML metadata

It is important that all of the pieces of the shapefile are kept together in the same folder, otherwise the file will not work - so be careful when moving files around! Renaming files is often problematic - if you rename one

you must rename all of them with the same name, otherwise they won't function together. You can easily rename batches of a file with the same name but different extensions if you are familiar with using the command line (i.e. Unix/Linux shell or DOS Command Prompt); it's less tedious than renaming them by hand in a GUI (like Windows Explorer).

Adding Data to a Map View

When you add map layers or data to a map view, you are technically not adding data to the window, i.e. copying the file and inserting it into the project. Rather, you are establishing a link between the GIS interface and the files, which exist independently from the software. When you use GIS software to change the symbolization of the layers (colors, outline, labels, etc) you are not modifying the data file itself; you are simply telling the software to display the layers in a certain way. The software is essentially a window for viewing the data files. The only way to change the data files themselves (their geometry or attributes) is within an editing mode which you must specifically launch.

Drawing Order

For much of the 20th century maps were created by taking individual layers on translucent mylar sheets and laying them over top of a paper base map. For example, an outline of the United States with boundaries of each state could serve as a paper base map, with individual mylar sheets layered on top that had rivers and cities. The order of the sheets determined which features appeared on top, covering up other features. GIS functions the same way; the order of the layers determines which appear on top. If you move a polygon layer with a solid fill (i.e. counties) over top of a point layer (i.e. of schools), you will not see the schools as the county layer is covering it up. In order to show both layers, you would have to move the school layer on top of the counties.

Alternatively, you could make the counties layer hollow by removing the fill, which would allow the school layer to be visible if it was on the bottom. This solution isn't ideal, as the boundary lines of the counties would partially cover a school if that school was located on or near a boundary. Typically, you would use a hollow fill for a polygon if you wanted to display it's boundaries on top of another polygon layer that has a fill.

Section III: Exploring the Map View

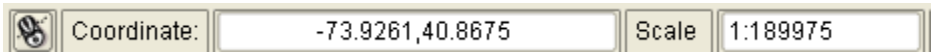
In this section you'll learn how to navigate the map view.

Steps

1. *Experiment with the Zoom tools.* Try each of the zoom tools in the Menu Bar.
   -  Zoom In - click to zoom in once, draw a box to zoom in to an area, or use the mouse wheel.
   -  Zoom Out - works the same as the Zoom In tool
   -  Zoom Full - will zoom the window to the maximum extent of all visible layers
   -  Zoom to Selection - zooms to selected features (skip this one for now)
   -  Zoom to Layer - zooms to the maximum extent of the feature currently selected in the ML
   -  Zoom last - returns to your previous zoom

- Zoom next - moves you forward to your next zoom (if you've already used zoom last)
- Refresh - redraws the screen (useful if your layers didn't draw completely or properly)
- Pan - move around the map by holding the left mouse button down and drag (does not change the zoom)

2. *Notice change in coordinates.* Move the cursor around the map. In the Status Bar (below the Map View) notice how the coordinates change; coordinates for the map are provided based on the position of the cursor. The unit of measurement is determined by the coordinate system and map projection of the project (since the project is in NAD83, the coordinates are in degrees and represent latitude and longitude). The scale box can also be used to change the zoom (a higher number to zoom out and a lower number to zoom in).



Commentary

Measuring Distances and Area

While QGIS does have tools for measuring distance, area, and angles, the utility of the tools is limited based on the current projection of the project. Since our project is using the coordinate system NAD83, the measurements will be in degrees.

There is a way to over ride this and get measurements in other units. You could go to Settings > Project Properties > Coordinate Reference Systems (CRS) tab, and check the box that says Enable on the fly CRS transformation. Now if you use the measurement tools the output will be in kilometers instead of degrees. To change the output to miles, if you go to Settings > Options > Map Units, you can change the default measurement unit from meters to feet, which will give you measurements in miles when using the tools.

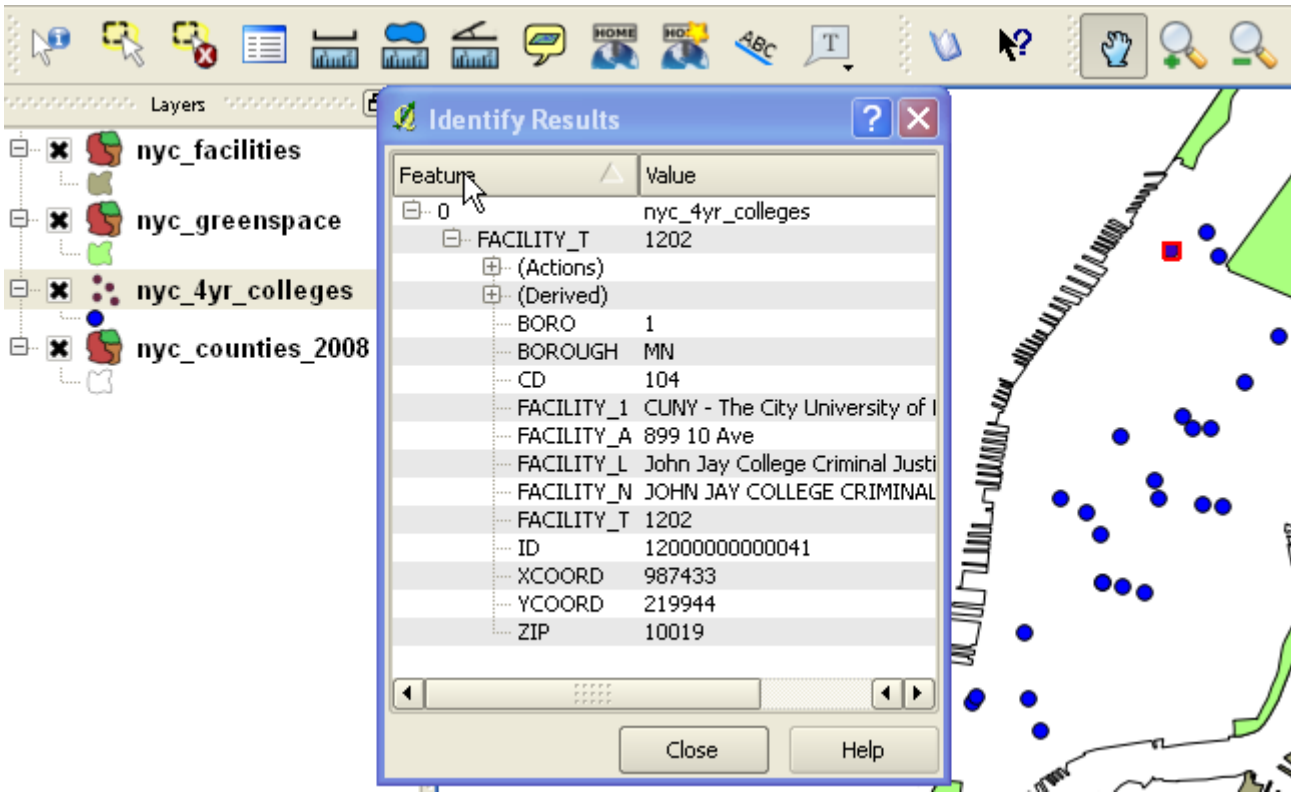
When you're finished measuring, it's a good idea to return to the Project Properties menu and change the settings back to the default (uncheck the on the fly box in the CRS tab and return distance measurements to degrees in the Map Units tab). If you forget to do this you may run into trouble later on (if you add additional layers that have different map projections or you try to create a new layer from existing ones).

Section IV: Exploring Features

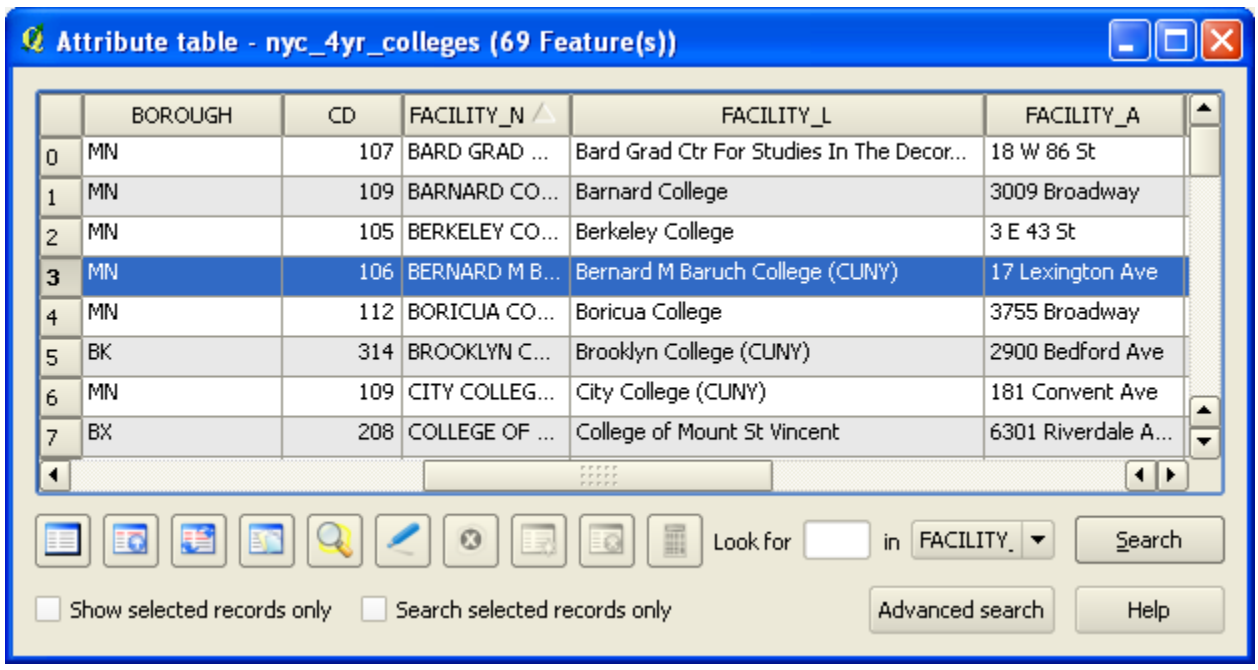
In this section you'll learn how to explore and interact with features in the Map View and Attribute table.

Steps

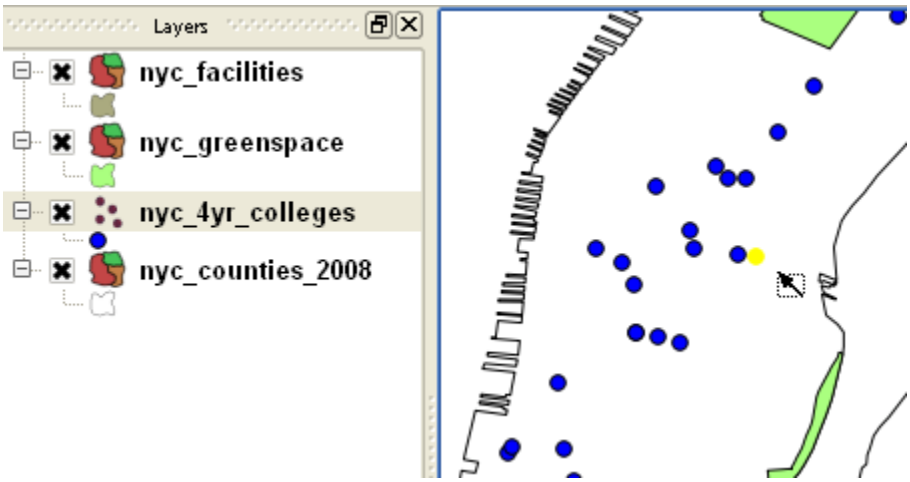
1. *Identify features.* Hit the Identify Features button in the Tool Bar. Select the counties layer in the ML. Click on Manhattan. Manhattan is hi-lited and information about that feature is displayed. Click on The Bronx to change the selection.
2. *Identify features from a different layer.* Make the colleges layer the active layer by selecting it in the ML. Click on any school in the map view to get information about that school. Where is this information coming from?

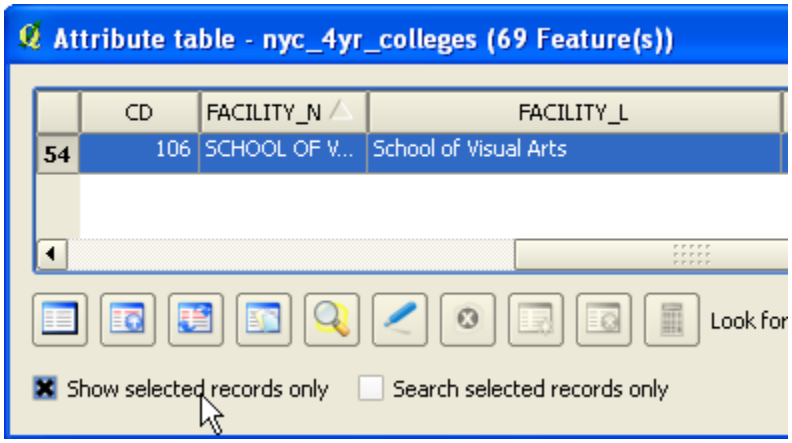


3. *Open the attribute table.* With the school layer still selected in the ML, right click on the layer and select Open attribute table (alternatively, you could click the  Open Attribute Table button on the toolbar). For every school (feature) in the school layer, there is a record for the school in the attribute table of that layer. Explore the table by scrolling across it and down.
4. *Select a feature from the table.* Sort the table by clicking on the field (column) heading that contains the name of the school (FACILITY_N). Click on the record for Bernard M Baruch College in the table. Close the attribute table. Zoom to the area around Baruch in lower Manhattan and you'll see it is selected. (Note - you can select multiple records from the table by holding down the CTRL key and selecting records one by one, or select a range by selecting a record, hold the SHIFT key, and select the last record).

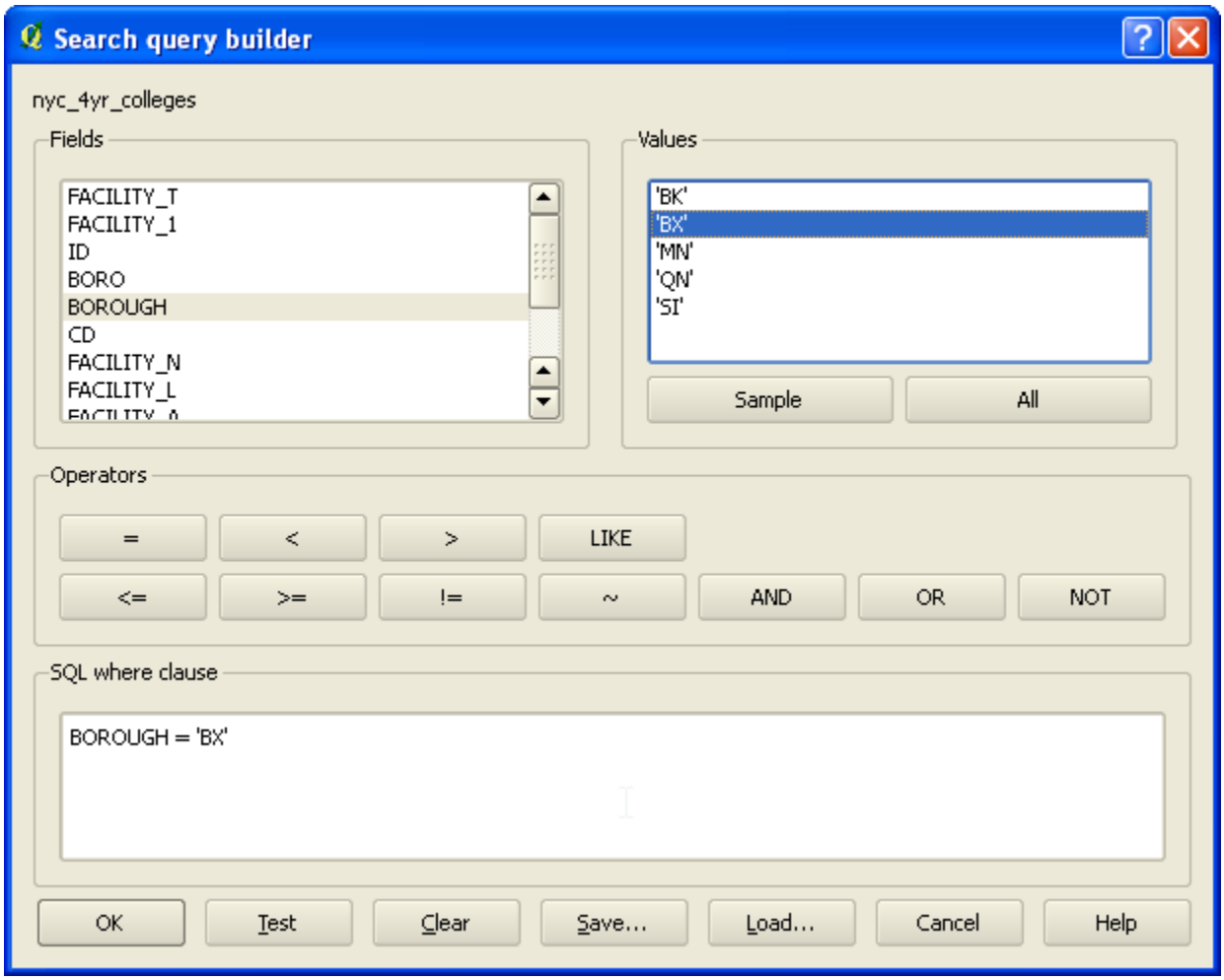


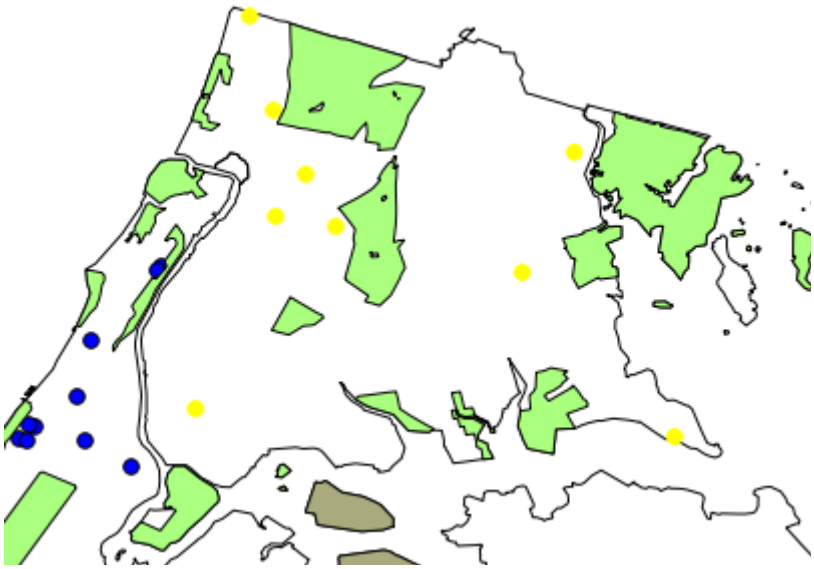
5. *Select a feature from the map.* Hit the  Select Feature button in the toolbar. Then select the school that is just to the east (right) of Baruch College. Hit the Open Attribute Table button. Click the checkbox that says Show Selected Records Only. This reveals the record for the School of the Visual Arts; this is the school that you've selected in the Map View. These two steps demonstrate that the table and map are linked, and you can select features in one and display them in the other. (Note - you can select multiple features by holding down the CTRL key and clicking on features one by one, or select several features by drawing a box around them). In version 1.6 the Select Feature button looks like this  and if you hold the mouse button down on the drop-down arrow it will reveal a menu with several select options: by rectangle, polygon, freehand, or radius.





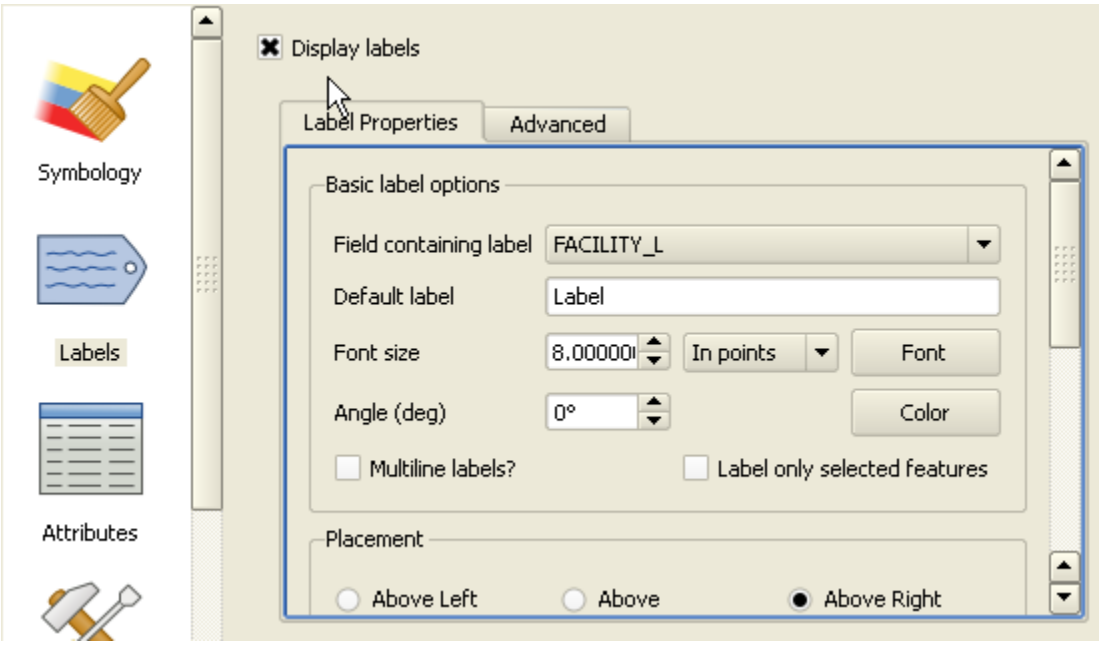
6. *Select Features by Attribute.* With the Attribute Table for the schools open, click the Advanced button in the lower right-hand corner. This opens the query builder window, which allows you to select features based on shared attributes. In the Fields box, double-click the BOROUGH field, which adds it to the SQL Clause box at the bottom. Click on the equals sign in the Operators section. Hit the All button under the Values box to display all of the unique values for the BOROUGH field. Double-click on the 'BX' value listed in the value field. Your statement in the SQL Clause box should read BOROUGH = 'BX'. Click OK. You've just selected all of the schools that are located in the Bronx. Close the attribute table and you'll see the schools selected in the map.





7. *Clear selected features.* Click the  Clear Selected Features button on the tool bar to remove selected features in the active layer (the active layer is the currently selected layer in the ML - in this case, the college layer). In version 1.6 the button looks like this:  . Alternatively, you could click on an area of the map that has no schools to clear the features, or you could clear the current selection from the attribute table.
8. *Labeling features.* Attributes stored in the table can also be used to label features. Double click on the school layer in the ML to open the Layer Properties. Go to the Labels tab. Check the box in the upper-left hand corner that says

Display Labels. In the dropdown box beside Field contains labels, choose FACILITY_L as the label field. Change the value in the Font drop down menu to font size to 8. Change the Placement radio button to Above Right. Hit OK. Explore the map a little. When you're finished, turn the labels off by returning to the labels tab in the properties menu for the layer and unchecking the box that says Display Labels. We'll experiment more with labeling later on.



Commentary

Attribute Tables

Every feature in the map view has a record in the attribute table; you can't have a feature without an attribute or vice versa. In a shapefile, the geometry is stored in the .shp file, an index of the geometry is in the .shx file, and the attributes are stored in a .dbf file. As we'll explore throughout this tutorial, attributes can be used for selecting, symbolizing, and labeling features in layers.

In GIS software attribute tables are managed and handled in the same manner as tables in a relational database. Each column has a data type associated with it which determines the kind of data that can be stored in that column and the types of operations that can be performed on it. Data types include strings (aka text) and various types of numeric fields (integers for whole numbers, reals for numbers with decimal places, etc). When you use the Query Builder to select features, like BOROUGH = 'BX', you are actually creating SQL code, which is a standard language for manipulating data in a database. The code 'BX' must be surrounded by quotes, as that is standard procedure when querying string (text) fields in SQL; if we were querying numeric values we would not use quotes.

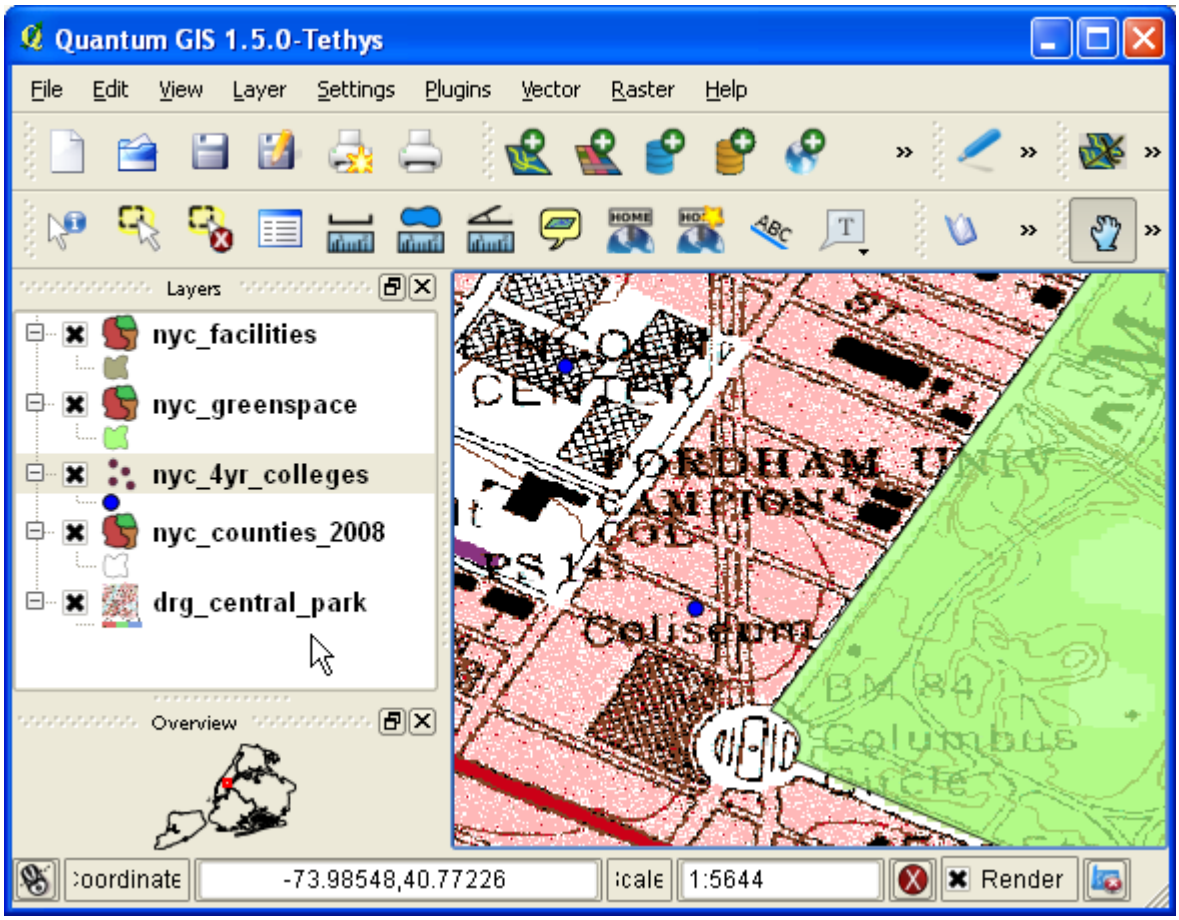
Section V: Adding Raster Data

In this section you'll get a very brief introduction to raster data.

Steps

1. *Add raster data.* Hit the  Add Raster Layer button on the toolbar. Browse to the data folder for part 2, select the drg_central_park.tif file and add hit open. Once the layer is added, drag it to the bottom of the ML.

2. *Explore raster layer.* Select the drg_central_park layer in the ML. Right click on the layer and select Zoom to best scale (100%). Explore the area of the map around Central Park and note how the raster layer lines up with the other layers. Select the parks layer in the Map Legend. Double click to open the Layer Properties and go to the Symbology tab. Drag the transparency slider to 25% and click OK. When you're finished exploring the map, uncheck the raster layer in the ML to turn it off and turn the transparency of the parks layer back to zero.



Commentary

Raster Data

Raster layers differ from vector layers in many ways including composition (continuous surface of pixels versus discrete geometric areas), file formats (many raster formats versus relatively few vector formats), resolution (optimal scale for raster layers matters more than vector layers), size (raster files tend to be much larger), and attribute tables (raster layers do not have attribute tables; the color of individual pixels denotes feature values). Given the differences in format, the tools for working with vector and raster layers are distinct (if you double click on the raster layer to open its properties, you'll see that most of the menu options are different from the vector layers).

Many geographic objects are represented in raster formats including satellite imagery, aerial photography, paper maps that have been scanned and digitized, and imagery that has been interpreted to represent value-added data that does not conform to political boundaries, such as land use and land cover and population density.

Up until quite recently the tools for working with rasters in QGIS have been limited, but this has changed with the addition of several plugins such as the gdal plugin, which allows you to perform raster analysis, and the georeferrencing plugin, which allows you to convert non-GIS image files (i.e. a scanned paper map) to a raster

GIS file by assigning coordinates to it. Given the time constraints of this tutorial, we're not going to cover rasters beyond this point. It was introduced here to give you a more complete picture of GIS capabilities and data formats.

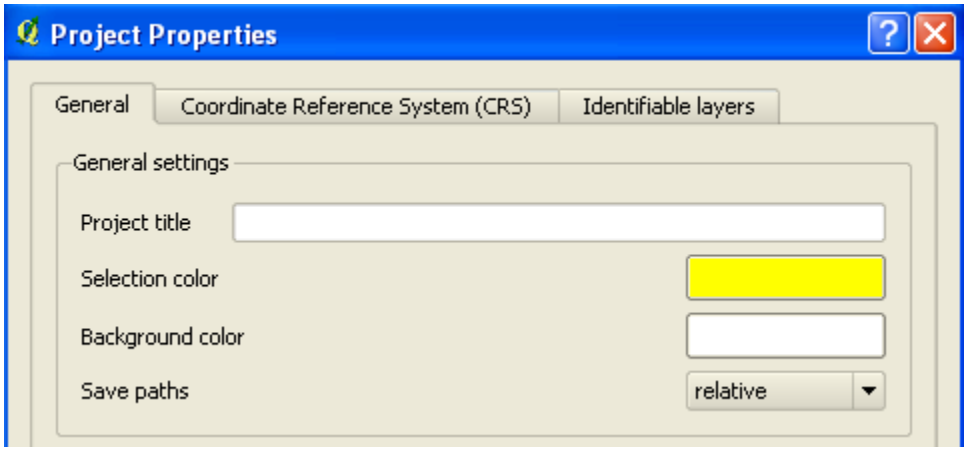
The raster used in this exercise is a DRG (digital raster graphic) which is a digitized, georeferenced version of the USGS' topographic maps. USGS topos are useful for studying elevation and terrain (particularly in non-urban areas) and for providing a frame of reference for overlaying vector layers or creating new ones; however most of the topos are several decades old and should be used with that fact in mind. The DRG was stored in a special .tif format called a GeoTIFF; a lossless image file that has georeferencing information (coordinates and map projection) embedded in it.

Section VI: Saving Your Project

You'll learn how to save your project.

Steps

1. *Change paths of files from absolute to relative.* Under Settings > Project Properties > General Tab, for the last option in the General Settings area labeled as Save Paths, change the drop down box item from Absolute Paths to Relative Paths.



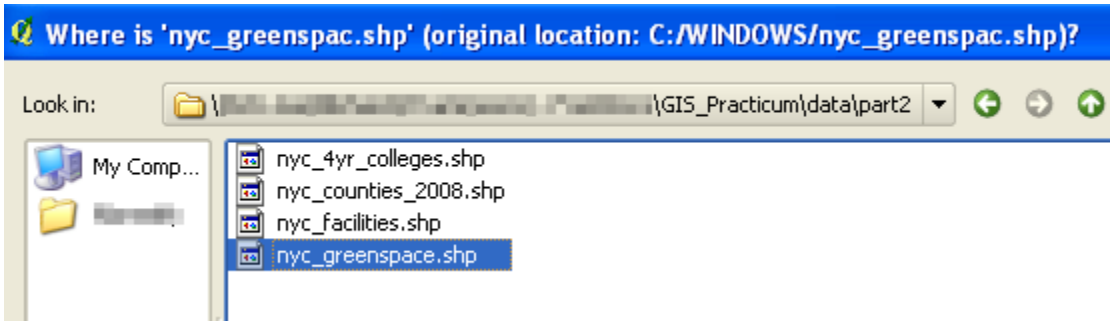
2. *Save your project.* Hit the  Save Project button. Navigate to the data folder for part 2, and save your project there as project2.qgs. The project file saves the symbolization, labelling, and current zoom for your data, and links to your data files (shapefiles); the shapefiles themselves are NOT stored inside your project file and exist independently. In order to use your project in the future, the project file and the shapefiles you used must be kept together.

Commentary

Project Files

When you add data to a project file you are not saving the data (shapefiles) inside the project; you are saving links to those files. Things like symbolization, data classification, the extent of your last zoom, and any finished maps you create are stored in the project file. When you click on the project file to open it, the software looks at the paths to your data, re-establishes the links, and then applies the settings (symbols, zoom, etc) that you have saved in your project file. This relationship is of crucial importance when it comes time to move or share files - if you move your project file or your data the links between them can become broken, and you'll need to re-establish the location between the project and the data in order to repair your project file.

If you open a project in QGIS and your project file can't find the data, because the data has been moved or renamed, the software will give you the opportunity to restore the link by asking you to browse through your file folders and select each file that corresponds to a layer you have in the ML of your project. Once you restore the links, you can save the project and it will save the new links.



Paths to files can be stored as absolute links or as relative links. An absolute link contains the complete path of a file, such as AS F:\My_Stuff\GIS_Practicum\part2\data\counties.shp. Use absolute paths when you're working in an established environment where you know that you won't need to move data and projects around, or in situations where your project files won't be stored directly above or in the same folder as your data. Absolute paths are a bad choice if you know you'll be moving data around; they're a particularly bad choice if you're working mobility on a usb drive in a Windows environment, as the paths can change as you move from machine to machine (i.e. F:\My_Stuff... on one machine becomes E:\My_Stuff... on another machine; QGIS won't be able to locate the files stored on F:\My_Stuff because it doesn't exist that way on the 2nd machine).

Relative paths save the directory and file information for the folder the project file is in (i.e. path would be .\county.shp) and all folders below it (i.e. path would be .\data\county.shp). Since anything above the project's directory is omitted, relative paths are a good choice if you know that you'll be sharing your project data or moving it around. Relative paths are a bad choice if your data is not going to be stored underneath your project folders (i.e. it's stored above the project directory, in a parallel directory, or another drive or server all together).

Think carefully about where to save project files in relation to your data, and once you've created your project file keep project files and data in a consistent place. Also remember that you must keep all of the individual components of the shapefile together (.shp, .shx, .dbf, .prj, etc); otherwise the shapefile will not function. If you want to share your project file with someone, you will also have to send them your data; the project file cannot exist independently from the data. You can share views or maps you've created in a static format (image file or PDF) that is separate from your project and data files; we'll explore that later in this tutorial.

The QGIS project file (.qgs) is actually just an XML file. If you open the project file in a text editor, you'll be able to see the structure of the file and all of its elements and attributes.

Part 3 - Geographic Analysis

The goal of part 3 is to introduce some analysis and geoprocessing features and techniques using a site selection problem as an example. Over the course of this exercise you'll learn how to: create a new project from an existing one, create a subset of a layer and process it to create land boundaries, join an attribute table to a shapefile, map the attributes of a shapefile, take a list of coordinates and convert it to a shapefile, draw buffers around a set of features, and select features based on their attributes and their spatial relationship to other features.

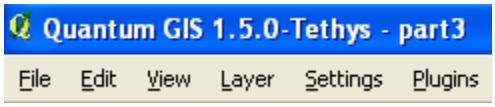
The object of this particular exercise is to identify potential areas within neighborhoods in New York City for locating a comic book store. Market research suggests that the primary demographic groups that purchase comic books are adults aged 18 to 34, people employed in professional and related occupations, and men. Based on this research we will identify neighborhoods that have a high percentage of adults in this age bracket and that don't have a large imbalance between the number of men and women. We will also identify areas within these neighborhoods that are within a half mile of a college or university (where young adults tend to congregate and people tend to work in professional occupations).

Section I: Creating New Project From Existing One

This section will show you how to create a new project from an existing one and will set the working environment for the rest of part 3.

Steps

1. *Open project.* Launch QGIS. Hit the  Open Project button (or go to File > Open Project). Browse through your folders to the QGIS project file you created for part 2, and select it to open it.
2. *Save Project As.* Once your project has loaded, hit the  Save Project As button (or File > Save Project As). Browse to the data folder for part 3. Save the project in that folder as part3.qgs. Hit Save. You've now saved a new copy of your old project, and are currently working in this new copy (you can tell by looking at the title at the top of the window, where the project name is listed). We will work with this new project, part3.qgs, for this part of the tutorial.



3. *Remove a layer.* We don't need the raster layer for this exercise. Select the drg_central_park layer it in the Map Legend (ML). Right click on the layer in the ML and select Remove (or, hit the  Remove Layer button on the toolbar).
4. *Zoom out and save.* Hit the  Zoom to Full Extent button to zoom out to the full extent of your layers. Then hit the

 Save button.

Commentary

Saving Projects and Removing Layers

Use the Save button to save the current project, and the Save As button to save the current project as a new copy with a different project name. Save As saves you the effort of starting from scratch if you have an existing project that you can use to branch off from. When you remove a layer from a project you're just severing the link between a particular project and that data; you're not actually deleting the data itself.

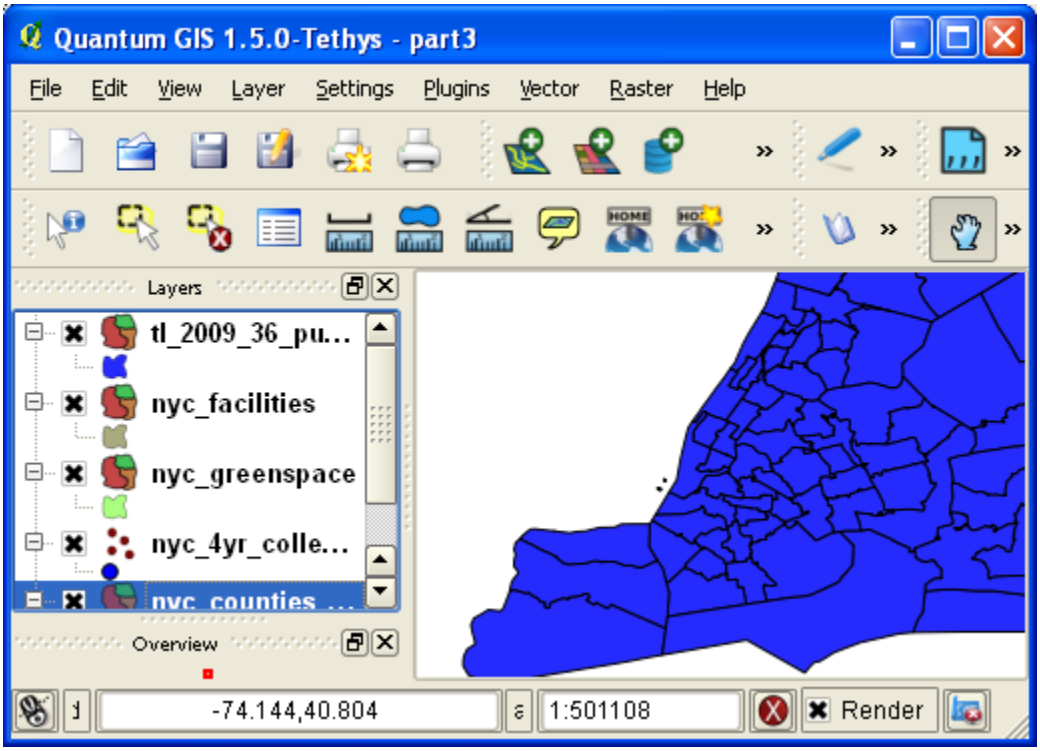
Section II: Geoprocessing Shapefiles

In this section you'll learn how to process a shapefile to prepare it for analysis. This is a common GIS task; normally when you download publicly available shapefiles you'll have to do some processing to make them usable for your projects.

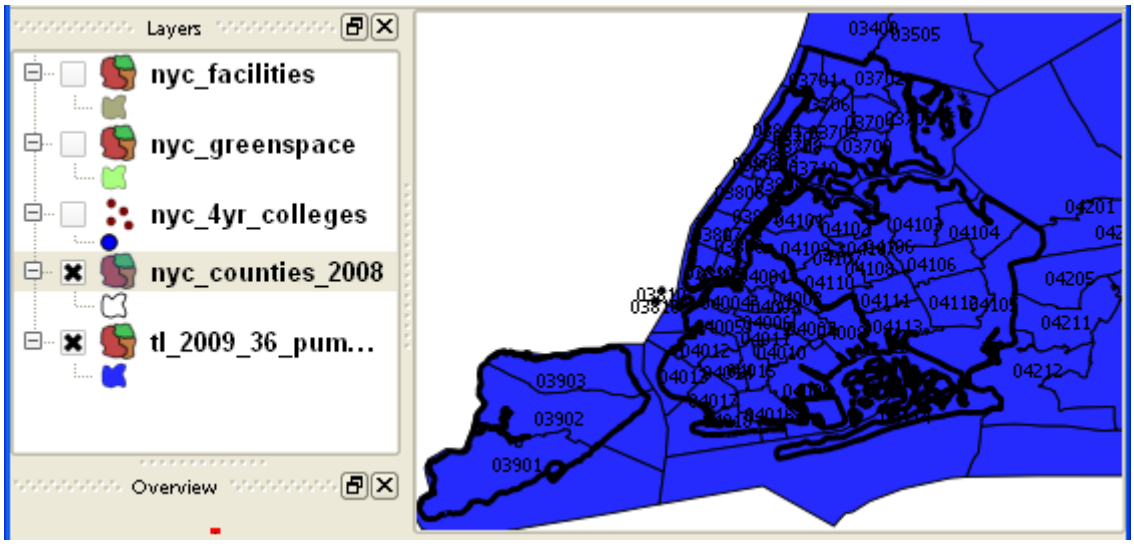
You'll be processing a boundary file for Public Use Microdata Areas (PUMAs) which we'll use to approximate neighborhood boundaries. PUMAs are statistical boundaries created by the US Census Bureau. The file was downloaded from the US Census TIGER Line Files.

Steps

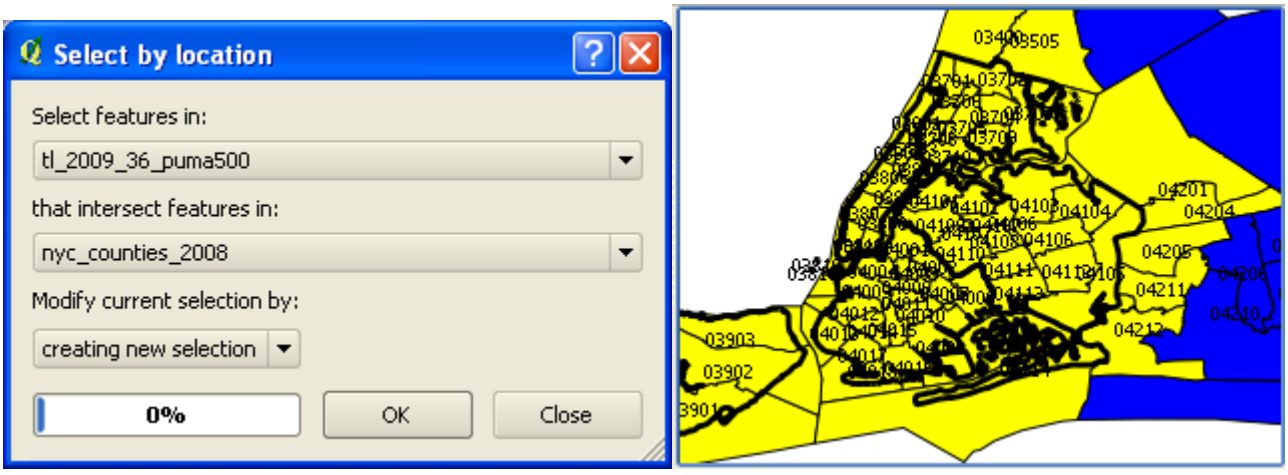
1. *Add the PUMA shapefile.* Hit the  Add Vector Data button. Hit the Browse button and browse to the data files for part 3. Select the PUMA layer, which is called tl_2009_36_puma500.shp, hit Open, and Open again to add the layer. By default the new layer will be drawn over top of the existing layers.



2. *Organize layers.* Select the PUMA layer in the Map Legend (ML) and hit the  Zoom to Layer button. You'll see the PUMA layer covers all of NY state, but we'll only need PUMAs for NYC. Select the counties layer in the ML and hit the  Zoom to Layer button. Select the PUMA layer in the ML, and drag it to the bottom of the ML. Check the boxes beside the green space, facilities, and colleges layer to turn them off for now.
3. *Change symbols for counties.* Double click on the counties layer. Under the symbology tab in the Outline option section, change the width box from .26 to .75. Then click on the labels tab. Check the Display labels box. In the Field Containing Labels dropdown, select the PUMA5CE00 field as the labels field. Change the font size to 8. Click OK to apply all the changes.

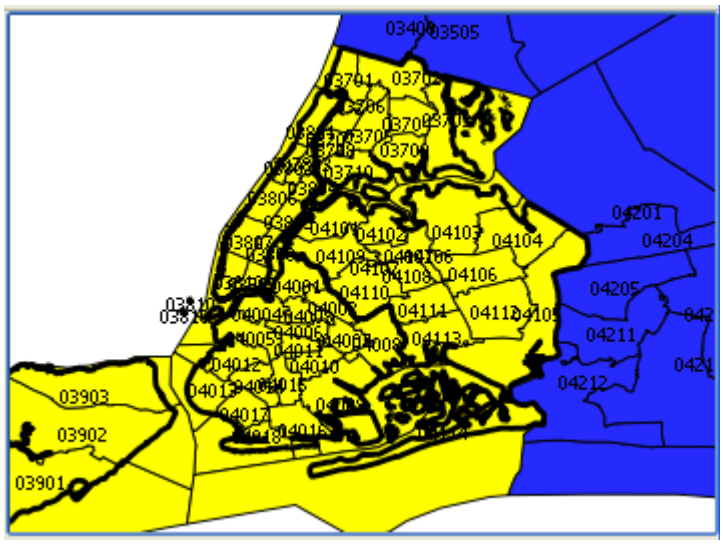


4. *Activate the fTools plugin.* If you haven't done so already, go to Plugins > Manage Plugins, and make sure the fTools plugin is checked. This will make the Vector menu appear on the menu bar.
5. *Select PUMAs within the counties layer.* Go to Vector > Research Tools > Select by Location. Select features in the puma layer (tl_2009_36_puma500) that intersect features in the counties layer (nyc_counties_2008), and keep the default for Modify current selection by creating new selection. Click OK. You'll see that all PUMAs within and touching the NYC county layer have been selected. Close the Select by Location menu when finished.

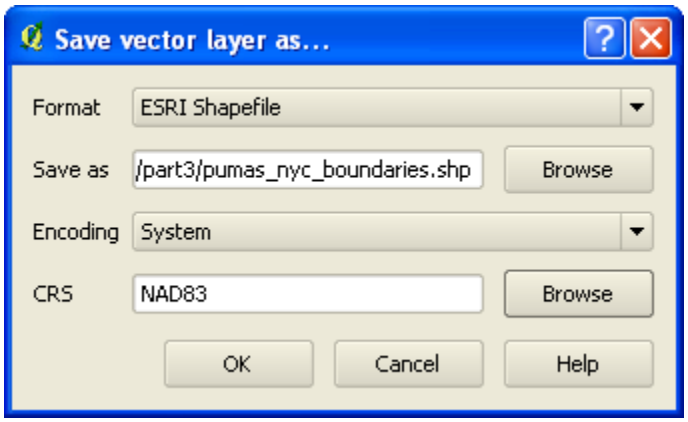


6. *Remove PUMAs outside NYC from selection.* Select the PUMA layer in the ML. Hit the  Select Features button (looks like this  in version 1.6). While holding down the CTRL key, click on each of the PUMAs that are outside

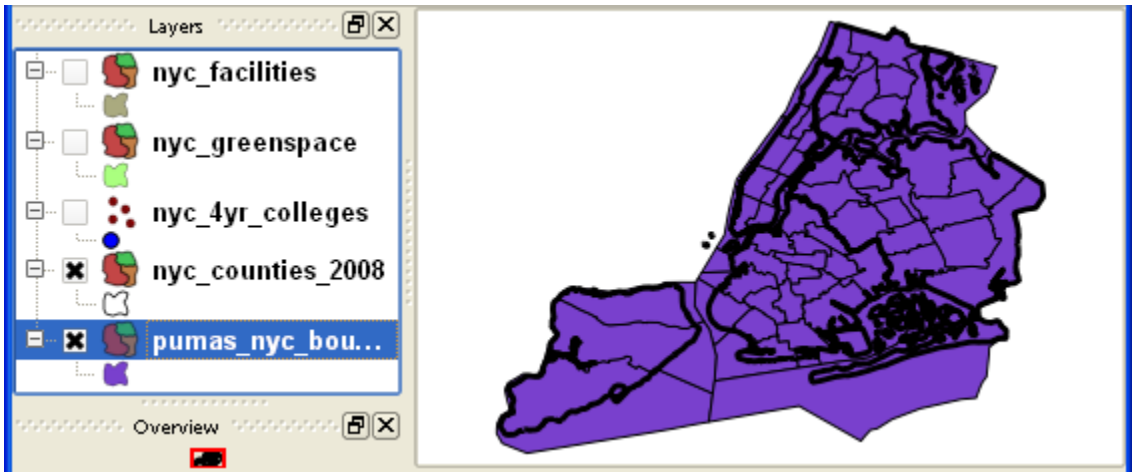
of the dark NYC boundary one by one to unselect each one. There are seven PUMAs that you must unselect (clockwise from top): 03400, 03505, 04201, 04204, 04205, 04211, 04212. If you unselect a PUMA by mistake, just click it again to re-select it. If you inadvertently unselect all of the PUMAs, you'll have to redo the previous step with the Select by Location tool to to reselect them.



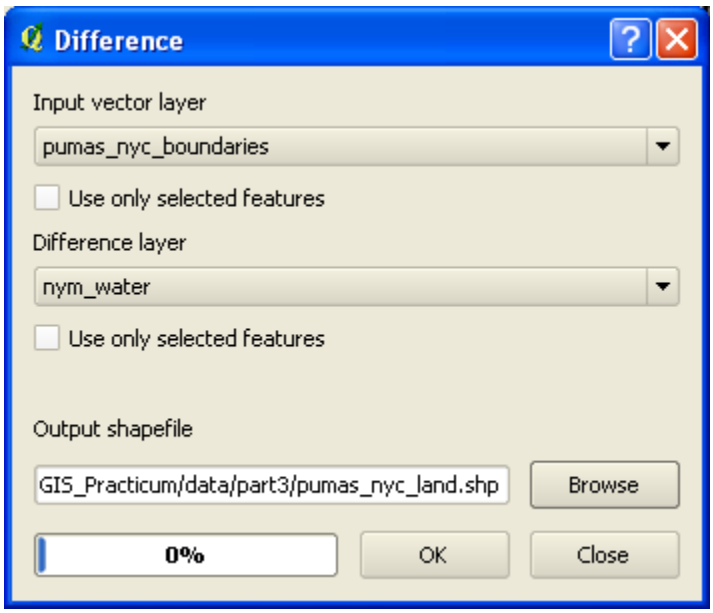
7. *Save selection as new layer.* Select the PUMA layer in the ML. Right click and choose the Save Selection As option. In the Save Selection as Menu, save the new layer as an ESRI shapefile. Browse and save it in your part 3 folder as pumas_nyc_boundaries. Leave the Encoding as the default, but hit the Browse button beside the CRS box and change the Original CRS to NAD 83 by selecting it from the list. Hit OK.

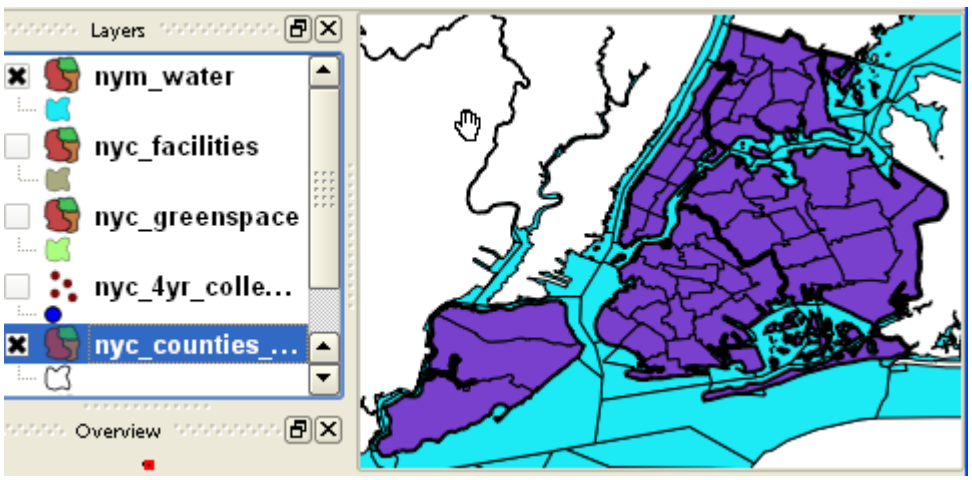


8. *Add new layer to map.* Hit the  Add Vector Data button. Browse to the folder to part3 and add the new layer you created, pumas_nyc_boundaries.shp. Drag it to the bottom of the ML. Then select the original PUMA layer for NY state in the ML. Right click and remove the layer, as we don't need it any more.  Save your project at this point.

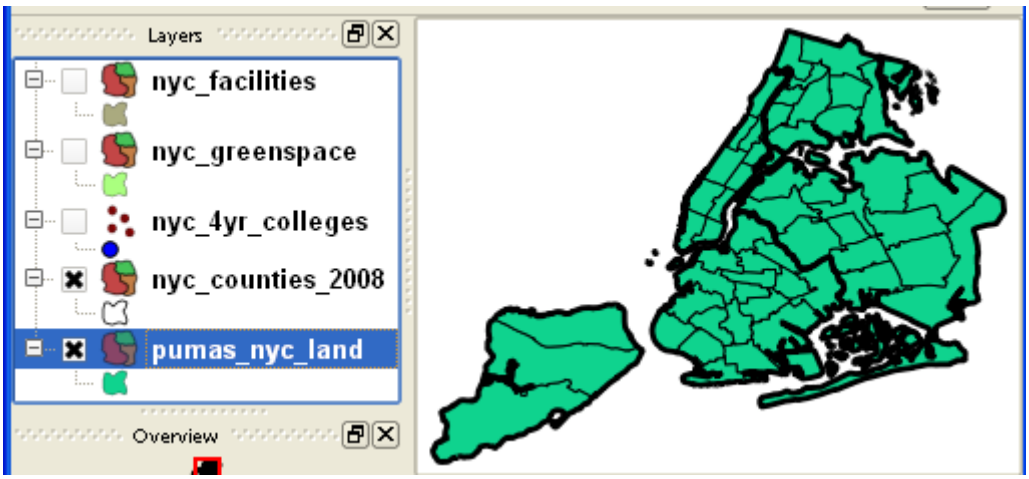


9. *Convert PUMAs from statistical boundaries to land boundaries.* Our last geoprocessing step is to convert the PUMA boundaries, which incorporate land and water, to boundaries that represent just land.  Add a vector layer, browse to the part 3 data folder and add the layer nym_water. On the menu bar go to Vector > Geoprocessing Tools > Difference. Select pumas_nyc_boundaries as the Input vector layer, nym_water as the Difference layer, and Browse and save the new file in your part 3 data folder as pumas_nyc_land. Hit OK. When prompted to add the layer to the project, say Yes. Close the difference menu.





10. *Clean up.* Select the nym_water layer in the ML, right click and remove it. Do the same for the puma_nyc_boundaries layer. Then drag the new nyc_pumas_land layer to the bottom of the ML. At this point, you have a brand new PUMA layer just for NYC that represents land boundaries. Save your project.



Commentary

Geographic Units

For this exercise we're working with Public Use Microdata Areas (PUMAs) which are a statistical area created by the US Census Bureau. While PUMAs were created for a specific purpose (geographically aggregating census microdata), they are also useful for mapping areas within large cities. PUMAs were designed to have approximately 100,000 people, which makes them better than legal or administrative units for making comparisons or mapping distributions. Neighborhoods in most North American cities are rarely formally delineated, so a geographic unit like a PUMA can serve as a proxy.

The choice of a geographic unit is an important decision; it's often a balance between the availability of data for an area, the suitability of the unit for the analysis, the amount of work that has to be invested in processing and analyzing the data, and the final outputs that will be created (tables, charts, maps) to explain the data.

PUMAs are a good choice for our exercise because: data is regularly published for these areas by the Census Bureau (annually as three year estimates published in the American Community Survey), PUMAs are good for approximating NYC neighborhoods, they are designed to have approximately the same number of people, and there are only 55 of them in the city which makes it manageable for this tutorial. One disadvantage is that the large size of a PUMA can mask individual population clusters within it, which makes it difficult to pinpoint an exact location for a retail store (but isn't unreasonable for getting a general idea of which areas to explore).

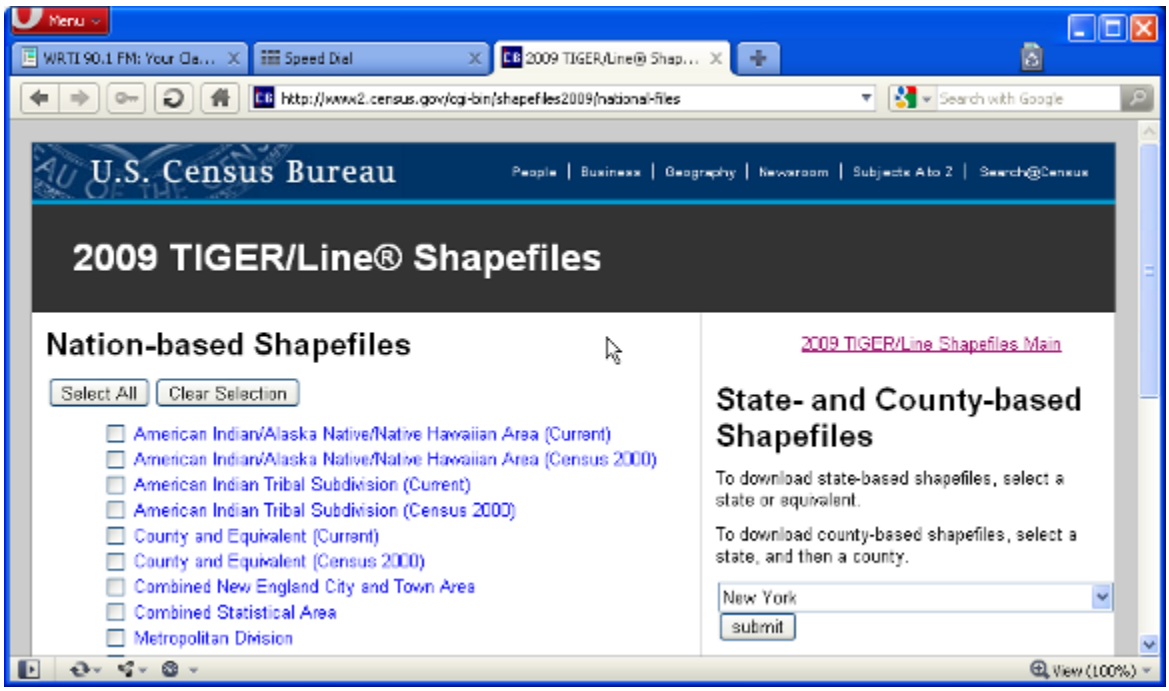
Compare PUMAs with other geographic units to get a better ideas of strengths and weaknesses. ZIP codes are more familiar to people and are commonly used in marketing, but the boundaries tend to be irregular and areas vary widely in size and population (ZIP codes were designed for delivering mail, not for studying populations). ZIP Code data is only available from the decennial census. Census tracts are census statistical areas created to have an optimum size of 4,000 people. They would be better for pinpointing a more specific location for a store, but there are thousands of them in the city and would require more time to process and work with. Annual Census Tract data is available from the American Community Survey in the form of five year estimates.

TIGER Line Files

The Census Bureau creates and maintains legal, statistical, and administrative boundaries for all geographic

areas that it publishes data for. It also creates and maintains geographic features such as water, roads, and landmarks that are used when creating statistical boundaries. These files were originally in a vector format created by the census called Topologically Integrated Geographic Encoding and Referencing or TIGER. The Census now provides this data in shapefile format. The files are in the public domain and can be downloaded for free at http://www.census.gov/geo/www/tiger/.

The PUMAs used in this tutorial were downloaded from the Census TIGER site. Most of the other files used in this exercise were created from the TIGER files. The NYC counties file is a subset of the TIGER county file for New York State, while the facilities and parks layers are aggregations and selections from the TIGER landmarks file for each of the five counties. All three layers were previously geoprocessed to convert legal boundaries to land boundaries, using a subset of the TIGER water features.



We were able to add the PUMA layer directly to our project because it shares the same geographic coordinate system as our other layers - NAD 83. Data downloaded from the Census TIGER site are all projected in NAD 83. We'll discuss and work with map projections later on in this tutorial.

Geographic Selection

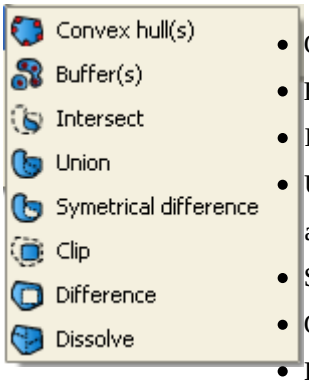
One of the strengths of GIS is the ability to perform spatial queries on features; i.e. select all areas that intersect other areas. This is one area where QGIS is still developing. The Select by Location feature of the fTools plugin only allows you to select features that intersect other features. However, several other spatial query options exist in other GIS packages, such as selecting features that border each other, or that are within or have their center within other features (the latter would have been the preferred option for selecting PUMAs within NYC counties). QGIS does have a Spatial Query plugin that can be activated in the plugins menu, and provides other options such as: crosses, disjoint, intersects, touches, and within. However, the tool isn't perfect and seems to have trouble when making selections between two polygon layers, which is why it wasn't demonstrated in this tutorial (although it works better when selecting points or lines in relation to polygons). If you need spatial query options beyond intersect, you can use other open source software: the command line GDAL / OGR tools, a geodatabase (PostGIS or SpatiaLite) tool, or GRASS GIS.

It's pretty common that you'll download geographic data that covers an area that is wider than you need. Since GIS data is malleable, it usually makes sense to grab data for a larger area and select out just the portions you need, if you can't find a layer that consists just of the areas you want; this is something to keep in mind when you search for data on the web.

Geoprocessing

It's also rather common that you'll download shapefiles that represent boundaries, but these boundaries will often incorporate land and water. If your intention is to show the actual boundary lines for reference purposes, then you will want to use the files as is. However, if you want to map the distribution of phenomena by area you'll want to process the boundaries to remove water as that phenomena isn't likely distributed there (i.e. there is no population living in the harbor or ocean). You'd also want to alter the boundaries if you're creating maps and want the user to be able to clearly understand the areas you're depicting. The Difference tool accomplishes this by subtracting the areas of bodies of water from the boundaries, resulting in features that show the outline of land.

This is merely one application and tool in the geoprocessing toolkit. Geoprocessing is essentially a GIS operation to manipulate the spatial aspects of GIS data. In the broad sense it includes layer overlay, feature selection, data conversion, and topology processing. In a more narrow sense that we're using here, it refers specifically to topology processing; modifying the actual geometry (points, lines, and areas) of features and files. Via the ftools plugin, QGIS has the following Geoprocessing tools for vector layers (running each tool creates a new layer; it does not modify existing layers):

- 

  - Convex Hulls - creates the smallest possible convex polygon enclosing a group of objects
  - Buffers - creates an equal zone around specific features at a specified distance
  - Intersect - creates new layer based on the area of overlap of two layers
  - Union - melds two layers together into one layer while preserving the features and attributes from both
  - Symetrical Difference - creates new layer based on areas of two layers that do not overlap
  - Clip - cuts a layer based on the boundaries of another layer
  - Difference - subtracts areas of one layer based on the overlap of another layer
- Dissolve - merges features within a single layer based on common attributes in the attribute table

In addition, there are also some geoprocessing tools under the Geometry Tools menu in ftools that convert or break polygons apart into simpler features like lines or points (we'll cover centroids, single-part and multi-part polygons later) and under the Data Management Tools menu (for aggregating many shapefiles into one file; the opposite of the selection / subset process). Geoprocessing for raster layers is available through the GDAL plugin.

File Naming Conventions

You may have noticed that when we've created new layers, we have used underscores instead of spaces when naming files, i.e. pumas_nyc_boundaries.shp. When naming files it's best practice to use underscores instead of spaces and to avoid using any punctuation in file names. This helps to insure compatibility of data across operating systems and to prevent possible errors when loading or reading data in the software. You should follow the same rules when creating folders to store data. The name of your file should reflect what it contains; you could include the geographic area it covers, the type of feature, and possibly a date or number to indicate different iterations of the data.

Section III: Joining and Mapping Attribute Data

In this section you'll learn how to join an attribute table to a shapefile and map the attributes in that table. Now that the PUMA boundaries are ready, we need to associate them with census data on the age and gender of residents of those PUMAs in order to select the optimal neighborhoods for locating our store.

Steps

1. *Open the data file.* Minimize (don't exit) QGIS for the moment. Using your file manager, browse to the data folder for part 3. Look for a file called acs_2006_2008_data.dbf. A dbf is a dbase file, used for storing data. This is a stand-alone dbf file that is not associated with a shapefile. Depending on what operating system you're using, open this file with a spreadsheet package like Excel or Calc (if you're in Windows, right click the file, select Open With, and then choose the option to select a program from the list. Choose Excel, hit OK, and open the file).
2. *Examine the data file.* The data file contains one row for each PUMA in NYC and several columns of attributes. The first five columns contain identifiers for each PUMA; the column ID2 is a FIPS code that we'll use to join this table to the shapefile. The remaining columns contain data from the American Community Survey. Columns are paired together, with the first representing the data itself and the second representing the margin of error (MOE) for the data. So, for the Riverdale / Kingsbridge PUMA (the first one in the spreadsheet), we're 90% confident (that's the confidence interval for the ACS) that there were 155,820 residents between 2006-2008, plus or minus 4,961. The columns that follow (with an associated MOE column) are: male population, percent of total population that is male, female population, percent of total population that is female, population aged 18-34, percentage of total population aged 18-34.

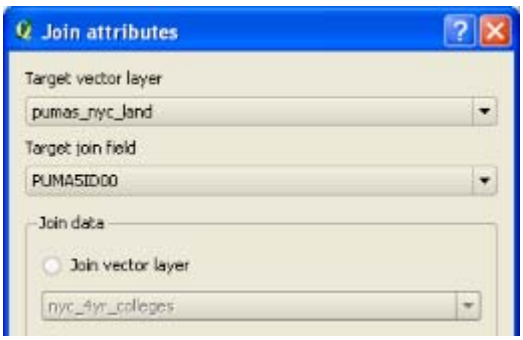
	B	C	D	E	F	G	H
1	ID2	GEO	BORO	NBHOOD	POP_TOTAL	MOE_TOTAL	MALE_POP
2	3603701	PUMA5 03701, New York	Bronx	Riverdale / Kingsbridge	115820	4961	53216
3	3603702	PUMA5 03702, New York	Bronx	Williamsbridge / Baychester	149185	5776	66780
4	3603703	PUMA5 03703, New York	Bronx	Throgs Neck / Co-op City	120081	4619	54359
5	3603704	PUMA5 03704, New York	Bronx	Pelham Parkway	128293	5251	60700
6	3603705	PUMA5 03705, New York	Bronx	Morrisania / East Tremont	153418	4805	70254
7	3603706	PUMA5 03706, New York	Bronx	Kingsbridge Heights / Mosholu	123216	5706	59018
8	3603707	PUMA5 03707, New York	Bronx	University Heights / Fordham	129521	5472	60640

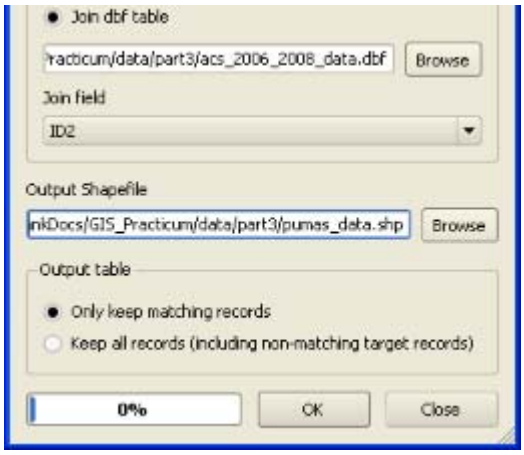
3. *Examine the attribute table of the PUMAs.* Close the dbf file, exit your spreadsheet software and maximize QGIS. Select the PUMA layer in the ML, right click and open the attribute table. In the table, note the column labeled PUMA5ID00. It contains the same FIPS code that was stored in the ID2 column in the data table: two digits representing the State of New York (36) followed by five digits representing the PUMA number. Since these columns are the same, we can use them to join the two files. Close the table.

Attribute table - pumas_data (55 Feature(s))

	STATEFP00	PUMA5CE00	PUMA5ID00	NAMELSAD00	MTFCC00
0	36	03701	3603701	PUMA 03701	G6120
1	36	03702	3603702	PUMA 03702	G6120
2	36	03703	3603703	PUMA 03703	G6120
3	36	03704	3603704	PUMA 03704	G6120
4	36	03705	3603705	PUMA 03705	G6120

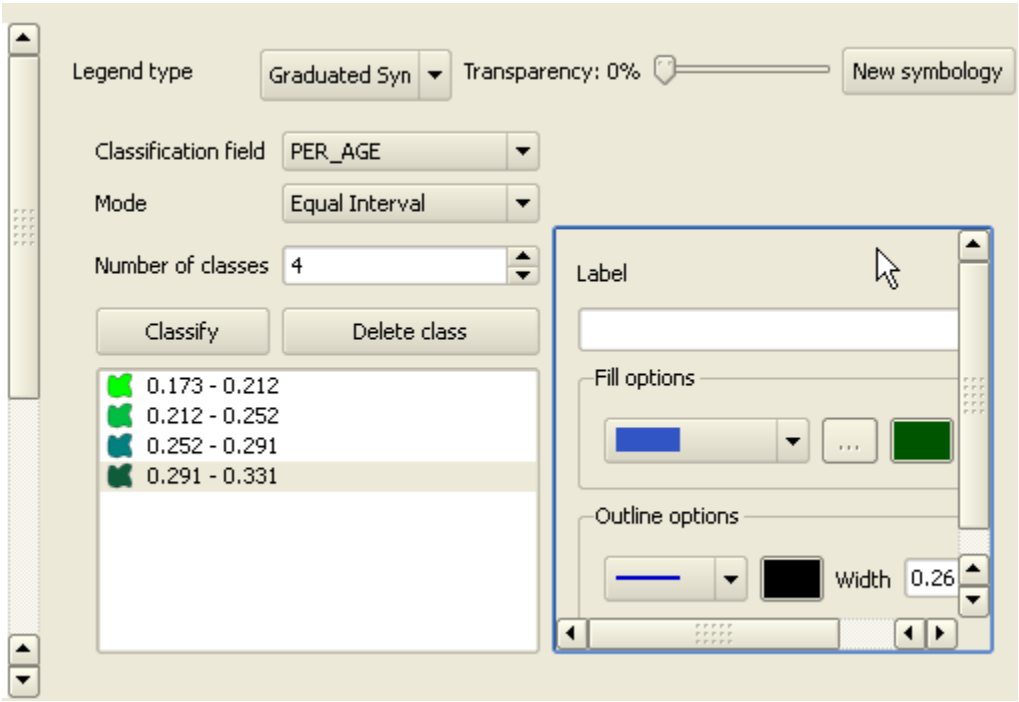
4. *Join data table to shapefile.* On the menu bar go to Vector > Data Management Tools > Join Attributes. Change the Target vector layer to pumas_nyc_land. Change the target join field to PUMA5ID00. Select the Join dbf table radio button. Hit browse to go to your data folder for part 3 and select 2006_2008_acs_data.dbf. Change the Join field to ID2. Hit browse to store a new shapefile (PUMAs with the data) in your part 3 folder - name it pumas_data. Keep the radio button that says to keep

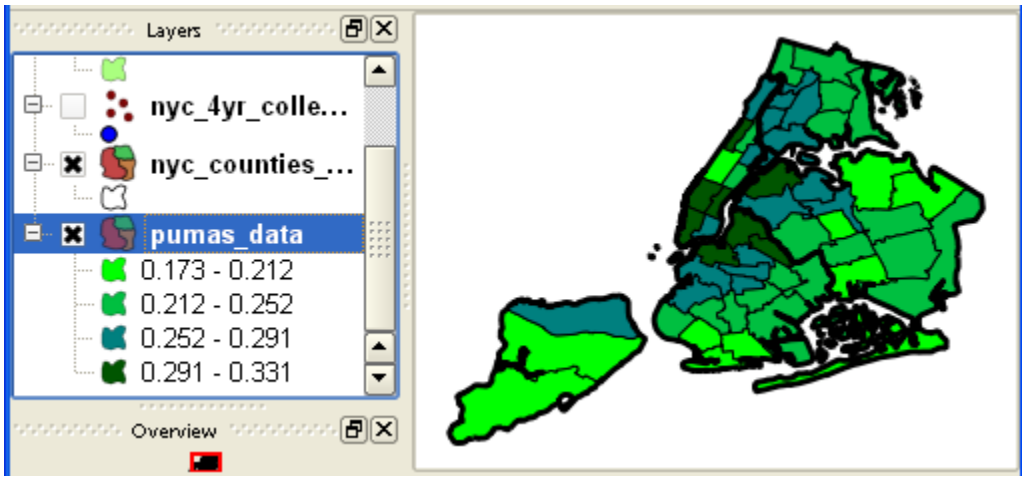




only matching records. Hit OK. Select Yes to add the new layer to the map. Close the join attributes menu. Right click on your new pumas_data layer in the ML and open the attribute table. Scroll over to the right, and you'll see that your new layer contains all of the data that was stored in the dbf file. Close the attribute table.

5. *Reorder the layers.* Select the pumas_nyc_land layer in the ML, right click and remove it. Drag the new pumas_data layer to the bottom of the ML. Save your project.
6. *Map the age data.* Double click the pumas_data layer in the ML and go to the symbology tab. Change the legend type dropdown from single symbol to graduated symbol. Change the classification field to PER_AGE (percent of population age 18-34). Keep the mode set to Equal Interval. Change the number of classes to four. Hit the Classify button. Select the last class in the Classify box. In the fill option, hit the fill box (which should be dark blue), change the color to dark green, click OK. Then click OK to apply all the symbol changes. You should now have a choropleth (shaded area) map that shows the percentage of the population of each PUMA aged 18-34, classified by equal intervals (divides data into categories that have an equal value range). Save your project.





Commentary

Census Data

The demographic data used in this exercise comes from the US Census Bureau's American Community Survey (ACS). Each year the census publishes annual survey data for all geographic areas in the US that have at least 65,000 people. Since the survey results for areas with smaller populations are often not statistically significant, the bureau averages data over several years for smaller areas. Data for areas that have at least 20,000 people is averaged for a three year period, and areas with less than 20,000 people down to the census tract level are averaged for a five year period. Each year the bureau releases a new annual data set and updates the two averaged data sets by adding the latest year of data and dropping the oldest one. For our exercise, we are using 3 year average data from 2006-2008. Even though PUMAs have a target population of 100,000 residents, it is better to use the three year data. As you drill down from the general population to figures that describe more specific groups, more data will be available in the three year dataset as the figures for smaller groups will not be significant in the annual dataset.

The American Community Survey was designed to provide data on a frequent basis and to largely replace the form on the decennial census that collected detailed information about the population. Beginning with the 2010 Census, the decennial census only provides basic demographic indicators of the population such as age, gender, race, and the total number of households and housing units. The decennial census is a count (not a survey) of the population and continues to be useful for making historical comparisons, providing a baseline for creating estimates, and for doing analysis below the census tract level (the decennial census is mandated by law to reapportion seats in the House of Congress). A third data product, Population Estimates, is published annually and is created using demographic calculations (as opposed to a count or survey) based on births, deaths, and migration. Basic estimates (total population, age, gender, race, and housing units) are published for states, counties, incorporated places, and metropolitan areas.

All the datasets from the US Census are available for download from the bureau's American Factfinder data portal at http://factfinder.census.gov/. All of the data is free and in the public domain. When you download the data you may have to process it to aggregate certain variables before you can use it. The age data that we are using in this exercise has been preprocessed; when initially downloaded, there was one column for each age cohort for each gender; the appropriate age and gender columns for the 18 to 34 population were combined and the unnecessary columns deleted.

Census data from other countries may be more difficult to obtain, as is may not be free or in the public domain, may not be documented in English, and may not be available in a digital format. You can check the website of the statistical agency for an individual country to see what is available, or you can visit the websites of international organizations like the United Nations or World Bank to obtain basic population data for all countries.

The decision of which census variables to examine in this study was made by consulting psychographic data and market research reports. This data is generated by marketing surveys to determine which groups of people are interested in products or activities relative to other groups based on age, gender, race, occupation, education level, and geographic location. The census data for this exercise was chosen based on statistics from the Market Reporter, a series of pyschographic reports published in a database called MRI+. This data is not freely or publicly available; you would have to access it through an organization that subscribes to the database, such as your university library, academic department, or place of work.

Identifiers

The ability to join data tables in a database or a data table to a shapefile is made possible by the use of identifiers, which are codes used to uniquely identify features. If features in two separate data tables share the same identifier, those data tables can be matched or joined together based on that common identifier, allowing you to create new data or to map data in a table.

There are several standard codes for identifying features. In the United States, FIPS (Federal Information Processing Standards) codes are a classification system for identifying all legal, administrative, and statistical areas in the country. For example, FIPS 36061 is the FIPS code for New York County (Manhattan). The first two digits are the code for New York State, while the last three digits are the unique code within New York State for New York County. In an attribute table these codes may appear in separate columns (state, county) or in a single column as one string.

The US government has also created two-letter alpha FIPS codes for each of the world's countries and uses them for international data published by various agencies. However, international data is more commonly coded with ISO codes (ISO 3166) which are available in a two-letter alpha format, a three letter alpha format, and a

three-digit numeric format.

Sample Country Codes

Country	FIPS 10	ISO 3166		
Denmark	DA	DK	DNK	208
Djibouti	DJ	DJ	DJI	262
Dominica	DO	DM	DMA	212
Dominican Republic	DR	DO	DOM	214

It is generally best practice to store ID codes as text and not as numbers since they don't represent quantities. Storing ID codes as numbers can result in data loss and misidentification. If codes begin with a value of zero and the ID is stored as a number, the zero will be dropped and the code will be incorrect (i.e. imagine you have a file with US ZIP codes and all ZIP codes that begin with zero are truncated).

In order to join two tables together based on an identifier, you need to be sure that each field is stored in the same data format; if one is stored as text and the other is numeric, the join will fail. Furthermore, you need to insure that each record is unique because one to many joins are not allowed; if you have a data table that has multiple records for one country, only one of those records will be joined to a shapefile and the others will be dropped. Finally, you should never use place names as identifiers or join fields because there are often many inconsistencies (imagine the number of different ways for spelling or abbreviating country names like the United States or South Korea).

Adding or appending identifiers to tabular data that lack this information is a common data processing task that you'll likely have to perform.

DBF Files

DBF files are an old data table file format for a database system called dBase, and were quickly adopted as a database table files for several database systems. While many of these databases are no longer widely used the file format has survived, in part because dbf files are a component of shapefiles that store all of the attributes of features. QGIS is able to take data stored in standalone dbf files and join them to dbfs affiliated with shapefiles based on a common ID code, using basic relational database techniques (a SQL join statement).

Important things to note about DBFs:

- You can view and create DBF files in spreadsheet programs such as any version of Open Office Calc and versions of Microsoft Excel between Office 97 and Office 2003. You can save text files and spreadsheets as DBF in these programs by using Save As and selecting DBF IV as the option.
- You can open or import DBF files with Microsoft Office 2007 and 2010, but you cannot save changes or create new DBF files because Microsoft decided to deprecate them. However, several plugins have been developed that allow you to work with DBFs in the newer versions of Office, and you can download these from the web.
- You can add dbf files or other data tables in QGIS 1.6 (but NOT previous versions) by adding a vector layer and selecting the dbf. This will allow you to view it, but to map it you'd still have to join it to a shapefile.
- DBF files are VERY particular - names for columns must be kept short (less than 10 characters), should contain no spaces or punctuation (except underscores), and cannot begin with numbers.
- Unlike plain text files, columns in a DBF table have a specific data type associated with them (text, integers, real numbers, etc). In order for joins between DBF files and shapefiles to work, the ID fields must be in the same format - text or numbers - IDs should normally be stored as text.
- You can open and edit DBF files that are associated with shapefiles. However - you should NEVER EVER re-sort the

data in a DBF file that is associated with a shapefile - if you do, the data will become misaligned with the features in the shapefile and will no longer match. You also CANNOT add new rows to the DBF, since there will be no geometry in the shapefile to match it. You can edit existing values, add new columns, and delete columns (as long as you don't delete the ID fields at the beginning of the sheet!)

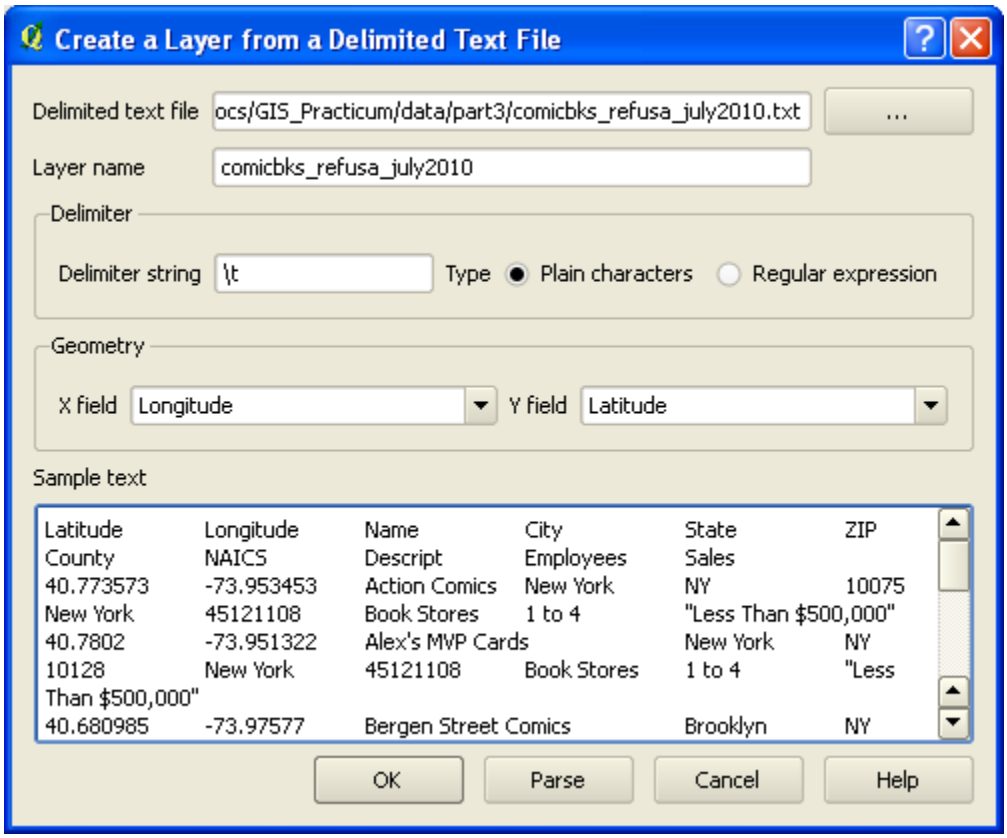
- If you need to do substantial editing of a stand-alone DBF file that is NOT part of a shapefile, it is best to copy all of the data in the dbf and paste it into a new, blank workbook in Excel or Calc format (xls or odt). For example, if you want to create a calculated field with percent change or do ANY work that involves formulas, create a new blank workbook - DO NOT work in the dbf file and do not create a second worksheet within the DBF - DBF can only support single worksheets. Once you finished doing the work in the spreadsheet file, do a copy and paste special in another workbook, pasting only values - no formulas or formatting. Then you can save that sheet as a new DBF file.

Section IV: Plotting Coordinate Data

In this section you'll learn how to take a text file with coordinate data, plot the data in GIS, and convert it to a shapefile. It's often difficult to find pre-existing shapefiles of buildings, particularly businesses and residences. But you can create your own point layers if you have the coordinates of the places you wish to plot. In this exercise you'll create a layer of comic book stores from a text file that lists each store with its latitude and longitude coordinates.

Steps

1. *Inspect the text file.* Go to your data folder for part 3, open the file comicbks_refusa_july2010.txt and examine it. This is a tab-delimited text file with data for comic books stores in NYC; each record represents one store and each attribute column is separated by a tab. Close the file when you're finished.
2. *Activate the delimited text plugin*. In QGIS, make sure that this plugin is activated under Plugins > Manage Plugins by checking it off in the list, and that the plugin toolbar is visible by right-clicking an empty area of the toolbar and checking the plugin box.
3. *Launch the delimited text plugin.* Click the  Add Delimited Text button or launch it from the Plugins menu. For the delimited text layer browse to the part 3 data folder and select comicbks_refusa_july2010.txt Accept the default layer name. For delimeter string, type in \t to indicate that the file is tab delimited. Hit the Parse button. Then make sure that the X field is Longitude and the Y field is Latitude. Hit OK.



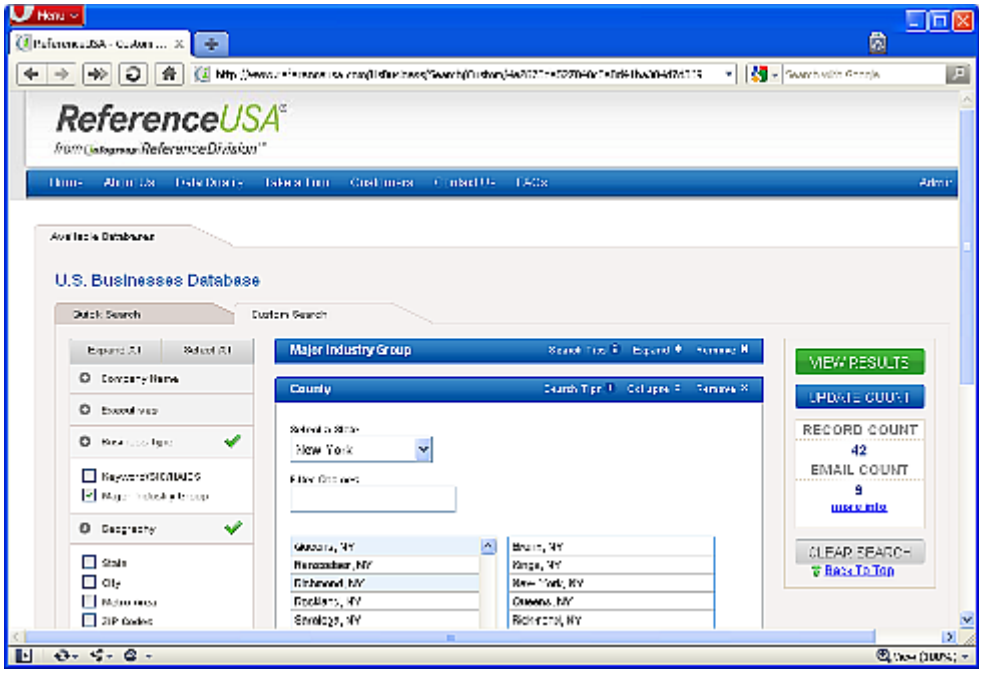
4. *Convert the plot to a shapefile.* Even though the points have been plotted, it isn't a shapefile yet. To convert it, right click on the comics layer in the ML and choose Save As. Save it as an ESRI shapefile in your part 3 data folder and call it comics_nyc. Change the default for the CRS to NAD 83.
5. *Add the new comic layer.* Hit the  add vector data button and add the new comics_nyc shapefile to your project. Then select the original text file in the ML, right click and remove it.  Save your project.

6. *View the attribute table.* Select the comics_nyc layer in the ML, right click and open the attribute table, to take a look at what's there. You should see all of the data that's affiliated with the comic book stores. Close the table when you're finished.

Commentary

Coordinate Data Sources

The coordinate data for the comic book stores was downloaded from a database called ReferenceUSA and processed so that it was ready for plotting. While government agencies often create and provide geographic data for boundaries and physical features, private features like businesses are usually not captured. These datasets must often be purchased or created from address or coordinate data. ReferenceUSA is not a freely available resource, but it is commonly held by many academic and public libraries. You can search for businesses by name, industrial classification code, and geography and download the data in spreadsheet format; although the number of records you can access in one download is limited. They provide comprehensive business, healthcare, and residence data for the US and Canada. The inclusion of XY coordinates (longitude and latitude) for each record makes it possible to plot the data in GIS.



There are free, public sources for downloading coordinate data that you can use to create features for natural (lakes, mountain peaks, parks, etc.) and human-made (cities, airports, schools, cemeteries, etc.) features, such as the USGS Geographic Names Information System (for US features) and the NGA's GEOnet Names Server (for international features). If you have batches of addresses, you can look-up the coordinates for these addresses using a geocoding service such as the Geocoding Service at the USC GIS lab.

Delimited Text Files

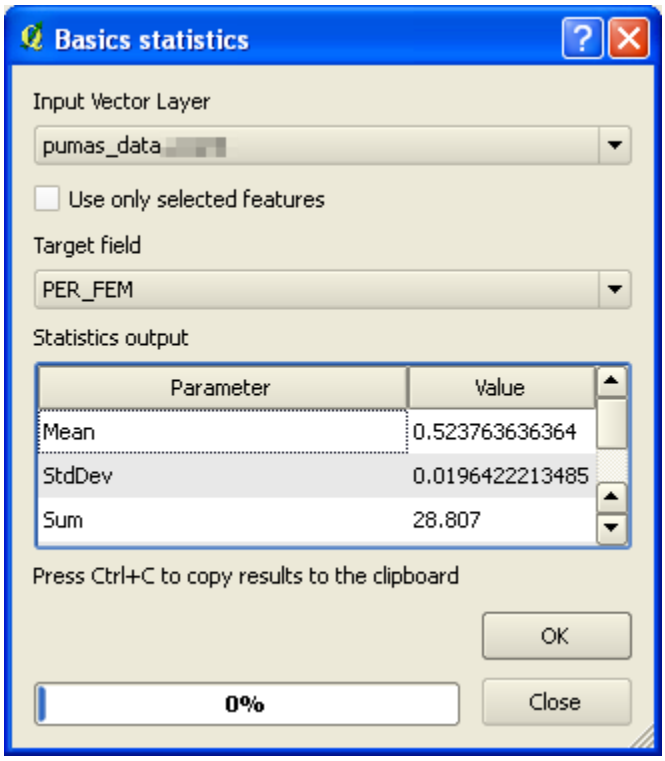
A text file is a plain document format that is often used for storing and sharing data. Since it is relatively simple and contains no formatting it is cross platform and historically stable. The attributes of each record are separated by a delimiter to indicate different fields. This allows spreadsheet and database programs to parse the text file into columns when you open or import it into that software. Common delimiters include commas, tabs, and pipes. The disadvantage of text files is that the fields are not associated with a specific data type; unlike a DBF file where a field can be designated as a string, integer, real, or other type. When importing text files you need to be careful that columns are designated correctly during the import process; strings inadvertently stored as numbers may have zeros dropped, while numbers inadvertently stored as strings cannot be treated mathematically. Depending on the source of the text files, fields that are intended to be strings may be surrounded by quotes, so that software can recognize and import those fields correctly. Storing data in tab or pipe delimited files is often safer than a comma file; if large numeric values have commas embedded in them (i.e. 5,000, 10,300, etc) the file can be parsed incorrectly as the software will assume that the commas represent new fields.

Section V: Running Statistics and Querying Attributes

In this section you'll learn to calculate basic statistics for attributes and use some of the advanced query features. Now that all of the data is in place, we can begin to remove neighborhoods that don't meet our site selection criteria. We want to target neighborhoods that don't have a large number of existing stores, have a high percentage of 18 to 34 year olds, and don't have a large imbalance between men and women.

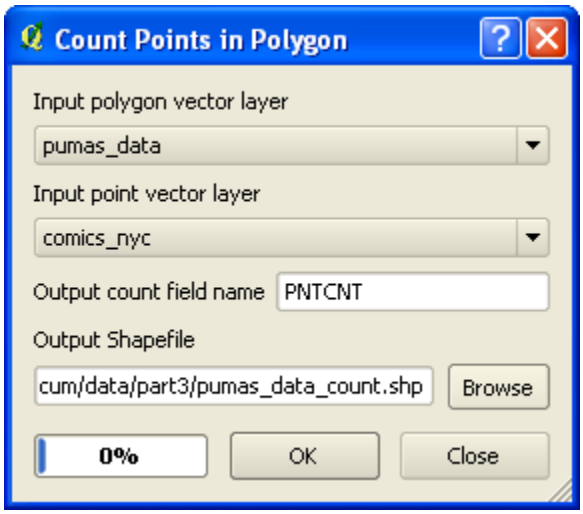
Steps

1. *Examine the age distribution.* By looking at the map we can see that, based on the two highest age categories (.252 to .291 and .291 to .331) the largest concentration of persons aged 18 to 34 is in lower and upper Manhattan, western Brooklyn and Queens, and the southern Bronx. But is this division of categories really significant? Open the attribute table for the pumas_data layer. Sort the data by the PER_AGE column. We can see the gap between the data class starting with .252 and the previous class is quite small; the previous class ends with .248. Furthermore, if you look at all the values from smallest to largest the distribution looks pretty consistent, with few sizeable gaps between values.
2. *Examine the gender distribution*. Sort the table by the PER_FEM column. You'll see that the PUMA with the highest concentration of women is 56%, and the one with the lowest concentration is approximately 48%. Overall, there really isn't a huge imbalance between men and women within each PUMA. Given this fact, for the purpose of our example we won't consider gender in our selection criteria.
3. *Run some basic statistics*. Close the attribute table. On the menu bar select Vector > Analysis Tools > Basic Statistics. Choose pumas_data as the input vector layer and the PER_AGE field as the target field. Hit OK. You'll see that the mean percentage is approximately .248 and if you scroll to the bottom, you'll see the median is .251. For the purpose of our example, we'll use 25% as our cut-off; PUMAS where 18 to 34 year olds make up 25% or more of the population will be included, while any with less than that number will be excluded. Close the stats menu.



4. *Count stores by neighborhood.* We should exclude PUMAs that already have a large number of comic book stores. On the menu bar go to Vector > Analysis Tools > Points in Polygons. Specify pumas_data as the Input polygon layer and comics_nyc as the Input point layer. Keep the output count field name as PNTCNT. Browse to your part 3 data

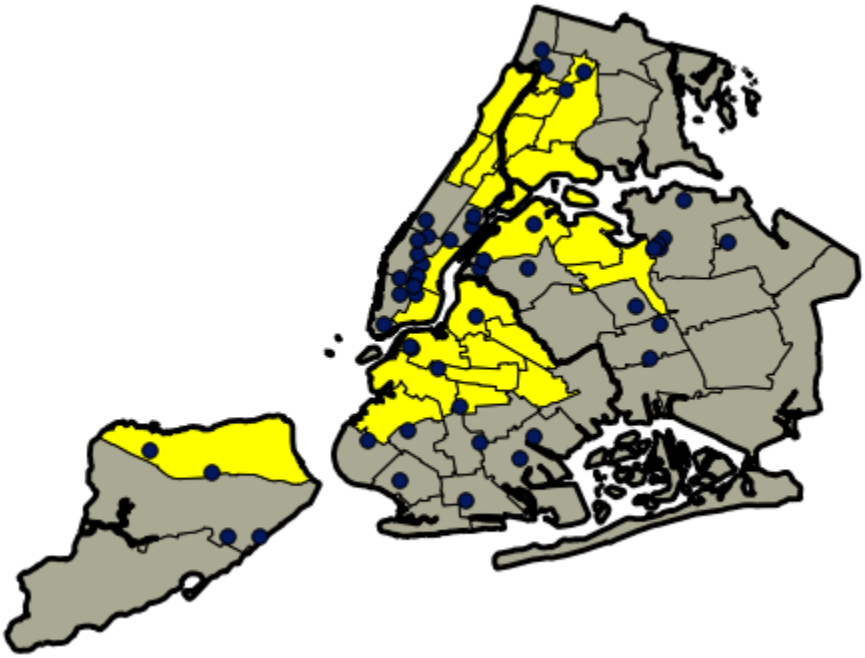
folder and save the output as pumas_data_count. Hit OK to create the new shapefile. Close the Point to Polygon menu.



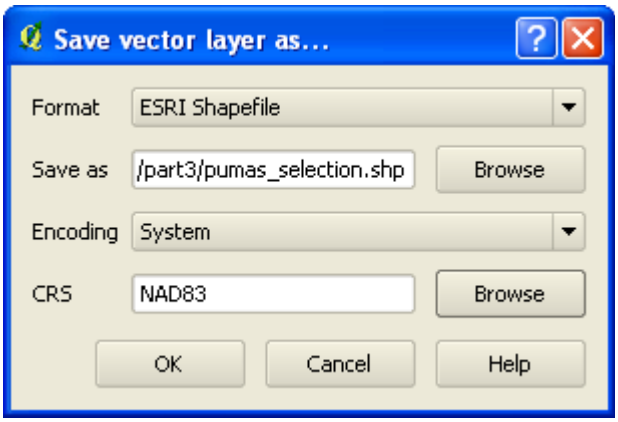
5. *Swap your layers.* Select the pumas_data layer in the ML, right click and remove it. Drag the new pumas_data_count layer to the bottom of the ML. Don't worry about symbolizing the new layer.
6. *View the table for the new layer.* Select pumas_data_count in the ML, right click and open the attribute table. Scroll the table all the way to the right. You'll see the new PNTCNT field, which shows the number of comic books stores in each PUMA. Click on the PNTCNT column heading to sort the table by that field. You'll see only five PUMAs have more than two comic book stores.

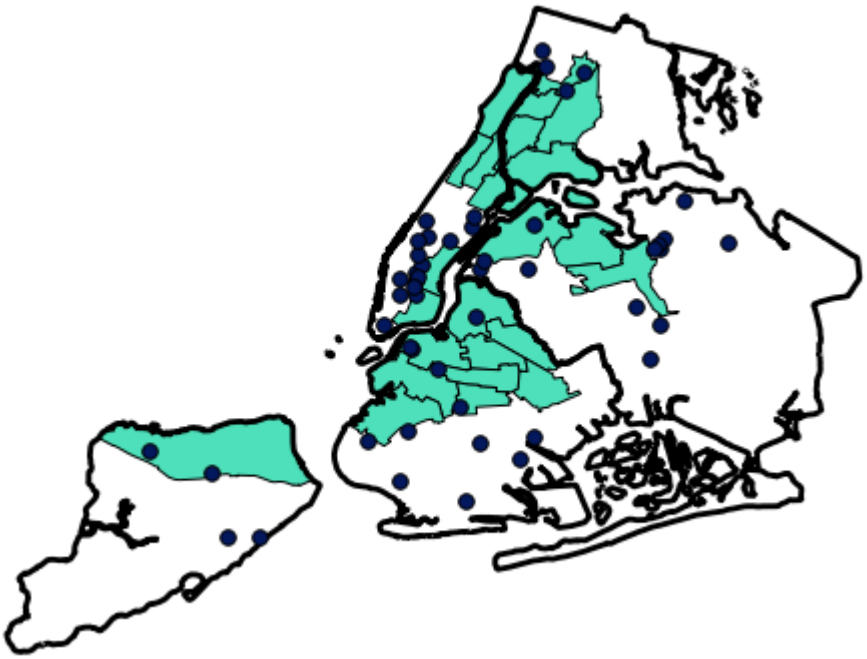
	RFEM	AGE_18_34	MOE_AGE	PER_AGE	MOE_PERAGE	PNTCNT
0	0.017	42916	2615	0.313	0.015	8
1	0.007	49286	2235	0.201	0.007	5
2	0.012	35864	2114	0.266	0.009	3
3	0.007	50204	2406	0.225	0.009	3
4	0.009	48742	2920	0.331	0.015	3
5	0.015	28407	2314	0.245	0.017	2

7. *Build an advanced query.* Hit the Advanced Search button in the lower right-hand corner of the attribute table menu. In the Fields box scroll down and click PER_AGE. In the Operators box click greater than or equal to (>=). In the SQL where to box type in the value .25 (be sure to include the decimal point). Hit the AND button in the Operators box. Double-click the PNTCNT field in the fields box. Hit the less than button (<) in the Operators box. Hit the All button under the Values box to populate it with a list of all PNTCNT values. Double-click on the value 3. In the SQL Where Clause box, your statement should read: PER_AGE >= .25 AND PNTCNT < 3. Hit the Test button to test your statement - you should have 24 features selected as a result. Hit OK. Close the attribute table to view your selections in the map view.



8. *Save your selection as a new shapefile.* Select pumas_data_count in the map legend (ML). Right click and choose the Save selection as option. Browse to your part 3 folder and save the selection as pumas_selected. Browse and change the CRS to NAD 83. Hit OK to save it. Hit the  add vector layer button to add the new pumas_selection layer to your map. Select the old pumas_data_count in the ML, right click and remove it. Drag the new pumas_selection layer to the bottom of the ML.  Save your project.





Commentary

Selection Criteria

Since the goal of our exercise is to demonstrate the capabilities and possible uses of GIS, we're not adhering to really strict criteria in our site selection process; the example is merely illustrative. Is a cut off of 25% of total residents aged 18 to 34 reasonable? It really depends on your goals, and whether you would prefer to have a focused, narrow selection of places or a more expansive one. Does it make sense to omit a PUMA that is only a few hundredths of a decimal place below 25%? These are the kinds of decisions you'll have to make for each project you do. You may decide that a line has to be drawn somewhere and that's it, or you may wish to allow an exception within a few decimal places or to round your numbers. You also could decide to make a qualitative decision - based on what you know about the neighborhood that's near the dividing line, should you include it or exclude it?

You have a few tools at your disposal for making these decisions; the basic statistics for determining mean, median, range, and standard deviation to establish a baseline are helpful. The data classification tools for symbolizing your data based on quantiles or equal intervals can also aid your decision (we'll discuss these later on). Regardless of what you do, look at the attribute table and make sure to examine and understand your data. You can easily copy the data from the attribute table using the copy to clipboard button and paste it into a spreadsheet, where you can create a scatter plot of the distribution to visualize where gaps in the data are. This can also aid you in classifying it (the natural breaks method; we'll discuss this later as well). It also helps to become familiar with the places you are studying, so you can draw on your more qualitative experiences to make decisions and perform a "reality check" on your observations.

Some Basic SQL

The advanced selection menu under the attribute table allows you to build complex queries for selecting features. QGIS, and most GIS packages, use the Standard Query Language (SQL) that's used when working with databases. Some tips:

- The boolean operator AND is exclusive; use it to select features that meet all of the criteria; the statement PER_AGE >= .25 AND PNTCNT < 3 will only select features where both criteria are met.

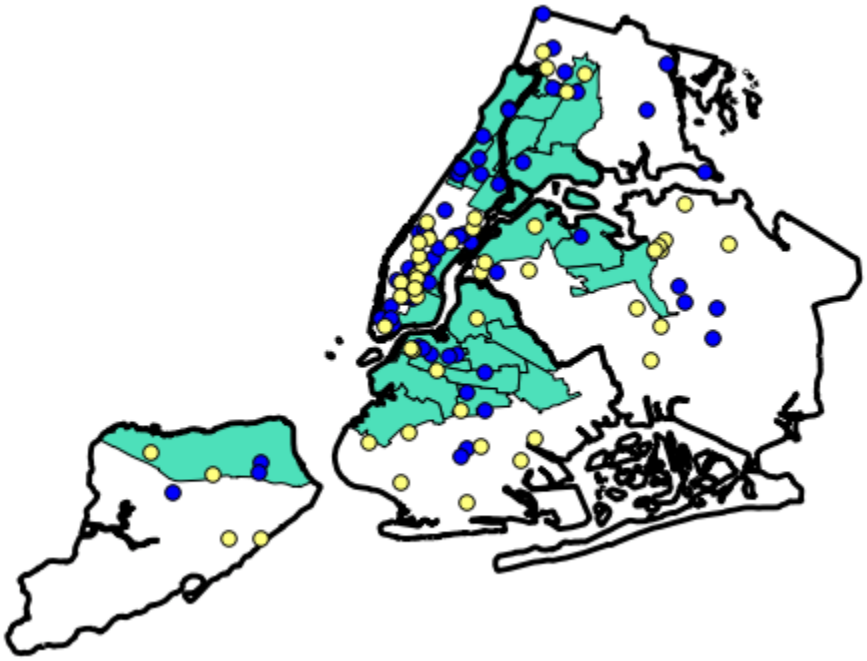
- The boolean operator OR is inclusive; use it to select features that meet one of the criteria; the statement PER_AGE >= .25 OR PNTCNT < 3 will select features that meet the first criteria, or the second one, or both.
- Your statements must be explicit; for every operation you must include the field that is part of the operation: PNTCNT > 3 AND PNTCNT < 5 is a correct statement. PNTCNT > 3 AND < 5 will yield an error, because you didn't specify the field for the second operator.
- Statements can be written more than one way. In our example above, which used only integers, PNTCNT < 3 and PNTCNT <= 2 yield the same result.
- If your query deals with text rather than integers, all text must be surrounded by 'quotes', otherwise you'll get an error. BORO='Bronx' will return all PUMAs in the Bronx. You can also use wildcards. BORO LIKE 'Bro*.*' will return all PUMAs in the Bronx and Brooklyn.

Section VI: Drawing Buffers and Making Selections

One of the primary strengths of GIS is the ability to layer different features and to combine or extract information to create new features. In this section you'll learn how to create buffers around features and to deduct areas from selections. For our example, since young people tend to congregate around universities we'll identify these zones and remove areas from our neighborhood selection that are not near schools.

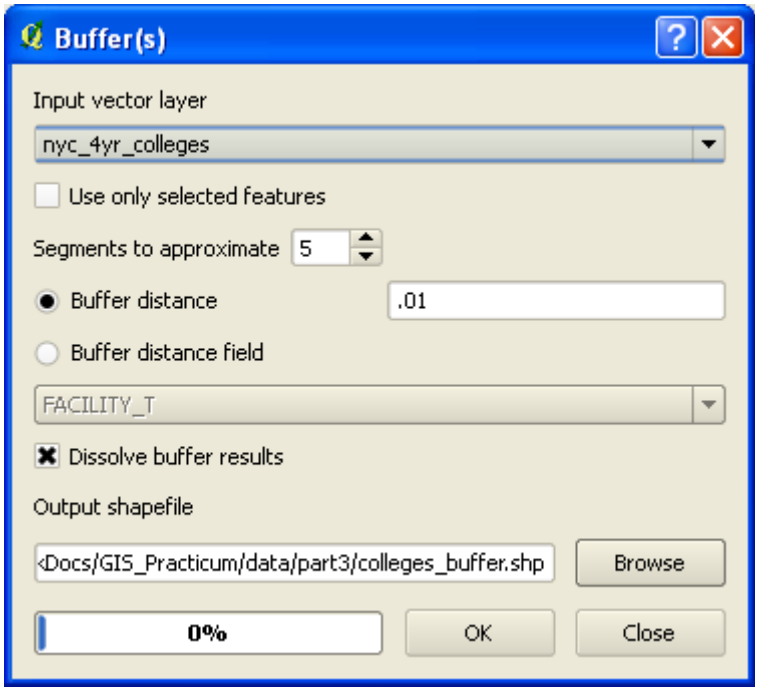
Steps

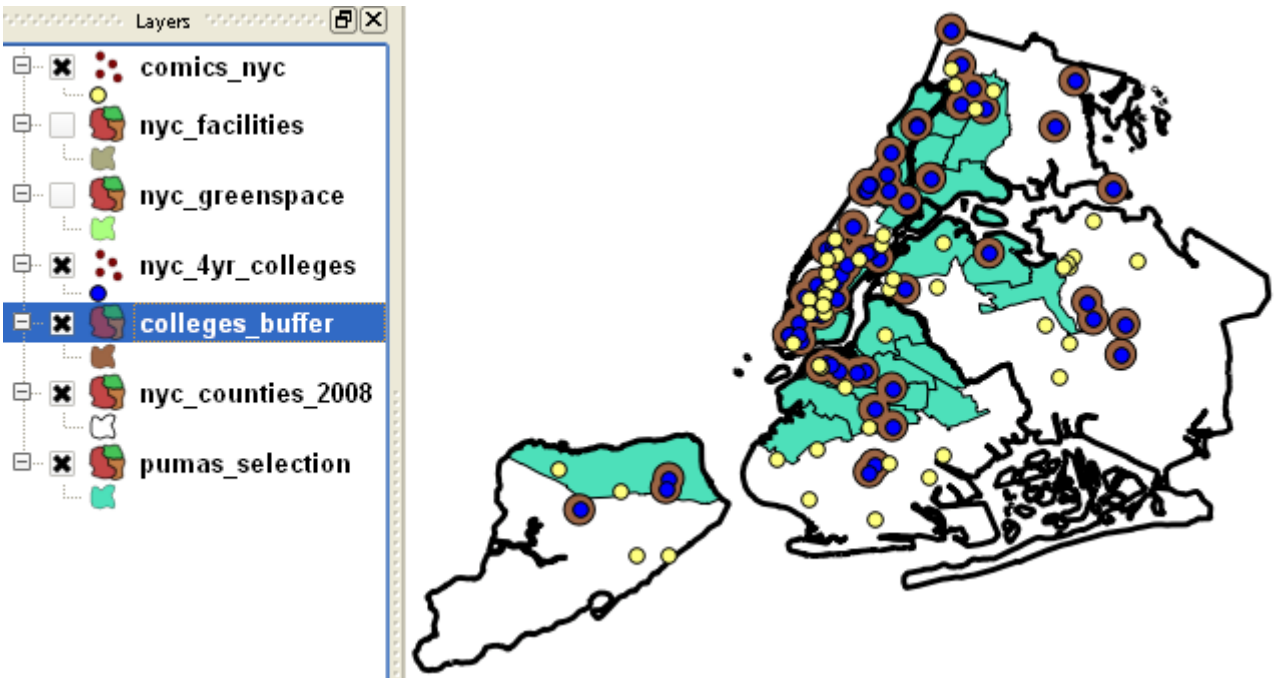
1. *Activate the colleges layer*. Hit the check box beside the nyc_4yr_colleges layer to turn it on. If it is similar in color to the comics layer, right click on one of the layers in the ML and change the fill in the symbology tab so you can clearly tell them apart (in our example, the comic stores are light yellow and the colleges are dark blue).



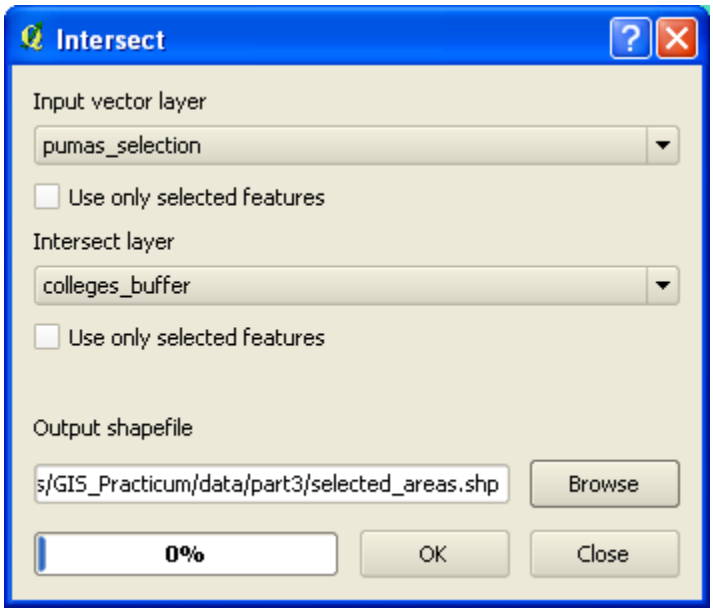
2. *Create buffers.* On the menu bar, go to Vector > Geoprocessing tools > Buffers. Specify the college layer, nyc_4yr_colleges as the input vector layer. For the buffer distance, type .01 (this is in degrees and represents approx 1/2 mile; see commentary below for explanation). Check the box that says Dissolve buffer results. Hit the browse button to save the new shapefile in your part 3 folder as colleges_buffer. Hit OK. Click Yes to add the new layer.

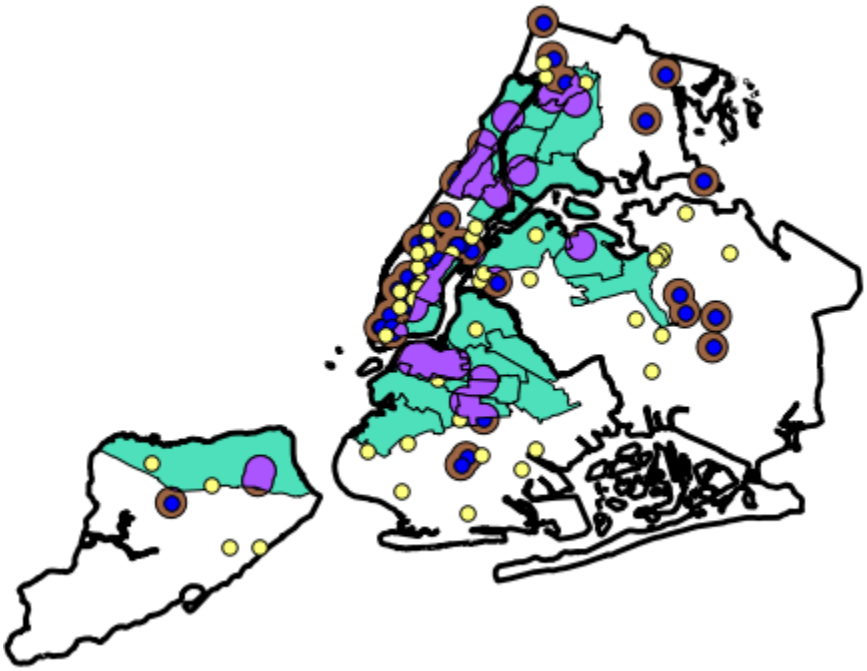
Close the buffer menu. Drag the buffer layer just below the colleges layer in the ML. Explore the map; you'll see a circular zone in a 1/2 mile radius around each college. The boundaries between each buffer zone are merged where zones intersect (as a result of checking the dissolve results box).



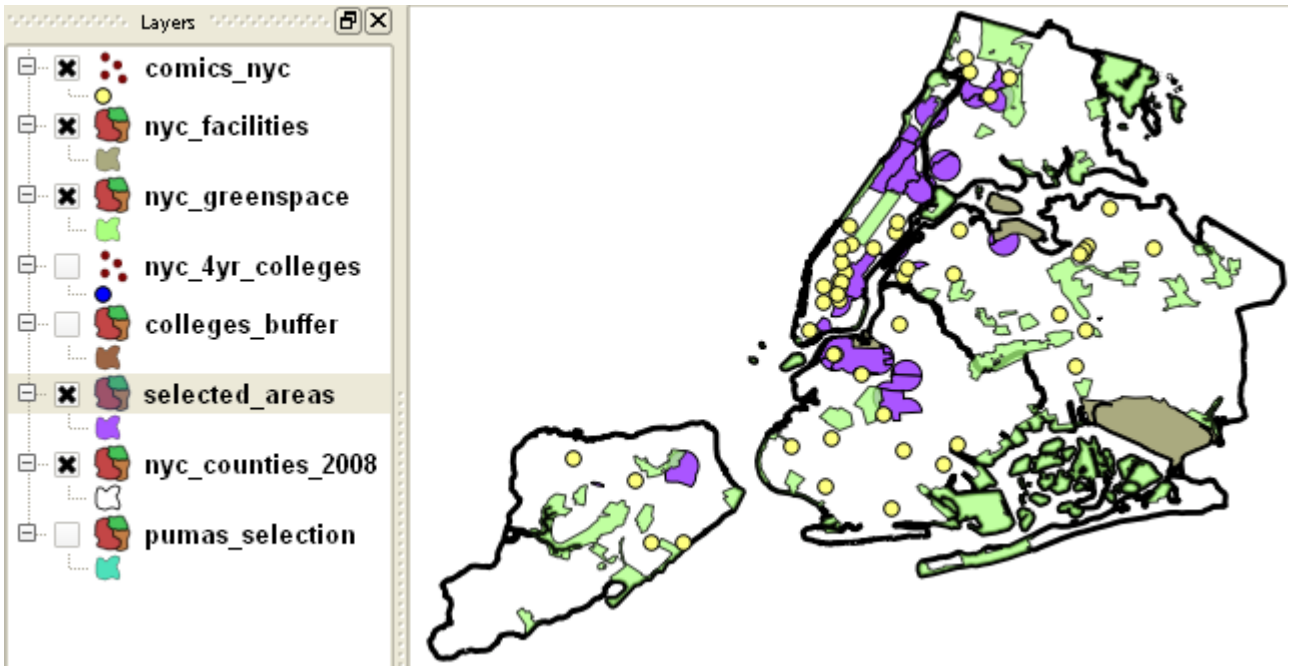


3. *Isolate areas within buffers and pumas.* On the menu bar, go to Vector > Geoprocessing tools > Intersect. Choose pumas_selection as the input vector layer. Choose colleges_buffer as the intersect layer. Browse and save the new result to your part 3 data folder as selected_areas. Hit OK. Close the Intersect menu. The new selected_areas layer shows you the areas to consider targeting: areas within a half mile of a college or university that are within PUMAS where the 18-34 age group represents 25% or more of the total population and there are less than three existing comic book stores.





4. *Clean up your map.* Uncheck the nyc_4yr_colleges, colleges_buffer, and pumas_selection layers in the ML to turn them off. Drag the areas_selected layer so that it is directly above the nyc_counties layer. Check the nyc_facilities and nyc_greenspace layer to turn them back on. We could refine our analysis a bit more by subtracting the green space and facilities areas that intersect our areas of interest, since we couldn't build a store on this land. For now, overlaying these land uses on top of our areas of interest should suffice.



5. *Identify areas.* Through the selection process, the attributes of our previous layers have been preserved in our new layers. Select the selected_areas layer in the TOC. Use the identify button and click on one of the areas. You'll see the attributes from our earlier PUMA layer. While the identifying information, like the name of the neighborhood, is useful, many of the other attributes are now incorrect. The population figures represent the entire PUMA and not the small subset we've selected. If we were going to save these layers for future analysis or projects, we would want to delete the attributes that are no longer necessary. Save your project.

Commentary

Buffers and Distance Measurement

One of the shortcomings of QGIS is it's inability to convert measurement units on the fly. Since the coordinate system of our layers is NAD 83 using degrees of latitude and longitude, we have to specify units for measuring distance in degrees. This is difficult for a number of reasons; it's much easier for us to conceive how large a kilometer or mile is relative to a degree. A thornier issue is that the length of a degree isn't constant - the distance between degrees of longitude decreases as we move from the equator to the poles. The distance between degrees of latitude is relatively consistent, but is also not equal to a degree of longitude, which requires us (or software) to make complex calculations to transform degrees into simple distance measurements. Here are some ways to get around this problem when creating buffers or using distance tools:

- Do some math and estimate. A degree of latitude at 40 degrees latitude is approximately 53 miles. If we want to draw a half mile buffer, divide 53 by .5 to get 106, then divide 1 degree by 106 to get approximately .01 degrees. So for our example a half mile is approximately .01 degrees.
- Use trial and error. Create a buffer in some unit of degrees, then measure the buffer using the measuring tool to see what it is in miles or kilometers. Then try drawing the buffer again, and base the number of degrees on your previous observation.
- Transform your layers to another coordinate system. Some coordinate systems use units like meters or feet instead of degrees, which would allow you to skip the unit conversion process all together. You can use the appropriate Universal Transverse Mercator (UTM) zone for the area you're studying, and the units will be in meters. In the

United States you can also use a State Plane system which is in feet. We'll cover projections and coordinate systems in the next part of this tutorial.

In our example we chose to dissolve the boundaries of the buffers where they intersected because we were interested in the total area within a half mile of any college. The resulting shapefile consisted of a single feature - the entire buffer. What if we wanted to preserve the individual boundaries of each buffer? We would leave that Dissolve box unchecked. The resulting shapefile would consist of several features, one buffer for each school, AND each feature would take the attributes of the school it surrounds (i.e. the school's ID codes, name, address, etc).

File Management

As we've moved through this exercise, we've created many shapefiles along the way; every time we made a selection or performed a geoprocessing function we ended up with a new file. There are two things we should note here.

First, this can get pretty confusing. With each new file you create, it's easy to lose track of what each one represents. You can mitigate this by giving your files names that clearly indicate what they are. Documenting your progress in a logbook, whether it's on paper or in a simple text file, can help you keep things straight. You may also decide to delete files that were created during the middle of the process. This is fine as long as you think you won't need to go back and re-do a step, either because the parameters of your project have changed or you've spotted an error.

Second, QGIS has recently made some improvements so it's not always necessary to create a new file with every single processing step. Some menus will give you the option to select features or perform operations on features that are ALREADY selected. This allows you to work with just the features you need from one layer to create a new one, skipping the interim step of creating a new shapefile of just the features you want to work with.

Site Selection

Site selection theories and land use analysis can be traced back to the early 19th century with the introduction of Von Thunen's land rent gradient. Subsequent work that included Weber's median location, Hotelling's competitive location problem, Christaller's Central Place Theory, and Tobler's Laws of Geography have provided a firm framework for the science (and art) of optimal site selection. Optimal site selection is studied within the fields of geography, location science, and operations management, and has expanded with the introduction and evolution of GIS. The three laws of location science, as explained by Church and Murray (2009) are:

- Some locations are better than others for a given purpose
- Spatial context can alter site efficiencies (the unique circumstances of a given area can alter whether or not a site is optimal)
- Sites of an optimal multisite pattern must be selected simultaneously rather than independently, one at a time (if you're planning to open several franchises you should do the planning all at once; as each site you open can impact another)

It's also important to understand the unique spatial patterns of each type of business or industry; a phenomena that economic and urban geographers have been studying for many decades. Products or services classified as low ordered goods tend to be located in most environments, and there will be more of these businesses in places with higher population densities. High order goods tend to require a higher population density and will be present in fewer locations. For example, businesses like gas stations, dry cleaners, and family

doctor's offices will be located in most areas, while office towers, specialty retail, and major hospitals will be located in fewer places, spaced further apart. Businesses like gas stations and convenience stores tend to cluster around major transportation intersections, while car dealerships and hotels tend to cluster around each other in districts. Movie theaters and large shopping malls on the other hand tend not to cluster together; they are spaced apart to serve different populations.

We worked with comic book stores in our exercise as they are good example of a specialized, high order business. They serve a highly defined demographic group and as an sub-industry they have not been co-opted by larger retailers or by the internet (at least not yet). Since there are approximately fifty in New York City our example was manageable. The use of PUMAs instead of census tracts was feasible; as a higher order business comic book stores attract customers from a wider geographic area relative to lower order businesses.

The location of non-retail or service industries is also distinct. Manufacturing industries often depend on the availability of raw materials and inputs and the distance for finished products to reach transportation and markets, while hi-tech industries tend to locate near pools of highly educated labor. Agricultural uses often appear where other land uses are not present and where land is inexpensive. The types of crops or livestock they produce will vary based on environmental factors like climate or soil.

The bottom line: if you are going to conduct a site selection analysis, you must understand the context: study the industry or business you are interested in, do some market research, make sure you're familiar with the geographic environment you're working with, and choose your geographic units of analysis and indicators carefully.

Section VII: Screen captures

In this brief section you'll learn how to create a screen shot of your map that you can easily share with others. You'll learn how to make a presentation quality map in the next part.

Steps

1. *Zoom to layer.* With the selected_areas layer selected in the ML, hit the  zoom to layer button. Use the  hand tool to center the map view.
2. *Save the map view screen.* On the menu bar, go to File > Save as Image. Browse to your data folder for part 3 and save the image there as map_screen. Change the Files of Type dropdown to PNG file. Click Save.
3. *View your map.*  Save your project and then close QGIS. Navigate to your data folder for part 3. Look for the file map_screen.png. Double-click it to open the file in your computer's default photo viewing program, and you'll see your map view. This is a quick way to save and share your map content. This is a simple, static image file that is not connected to your project or data files. You can easily email or text this file to anyone.

Commentary

Considerations and Next Steps

Based on our results, what would you do next? How would you decide where to locate the store? What else would you investigate? Is there anything that we've done in this exercise that you would do differently, if you had to conduct an analysis like this for an actual project?

For more practice, some things to try:

- Expand the selection areas to include more neighborhoods based on the percentage of the population aged 18-34.
- Instead of excluding neighborhoods that have 3 or more comic book stores, try excluding areas that are within a mile of existing comic stores (hint - a mile is equal to .019 degrees).
- Shrink the selection areas by removing the greenspace and facilities from the final areas.

Part 4 - Thematic Mapping

The goal of part 4 is to introduce you to map layout and design, as well as to some additional data processing techniques. You will also grapple with coordinate systems and map projections, which are central components underlying GIS. You'll learn about cartographic representation and design and the practical implications of choosing how to classify and represent your data.

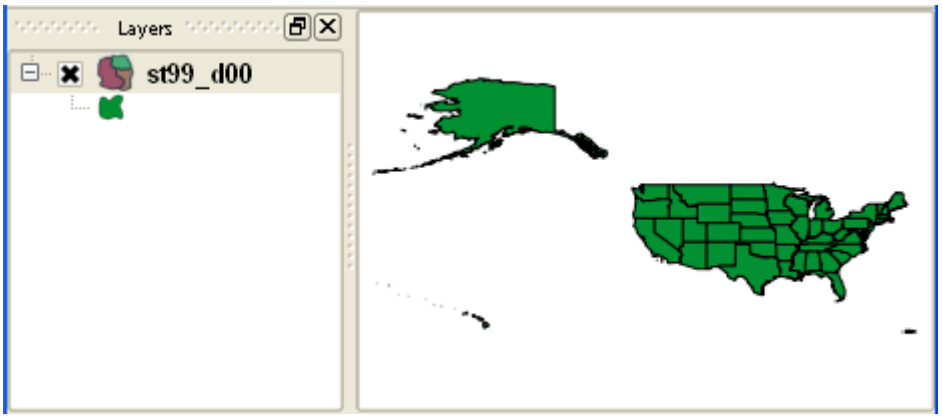
The goal of this particular exercise is to create a stand-alone thematic map to show the distribution of employment in the information sector by state in the United States.

Section I: Defining and Transforming Projections

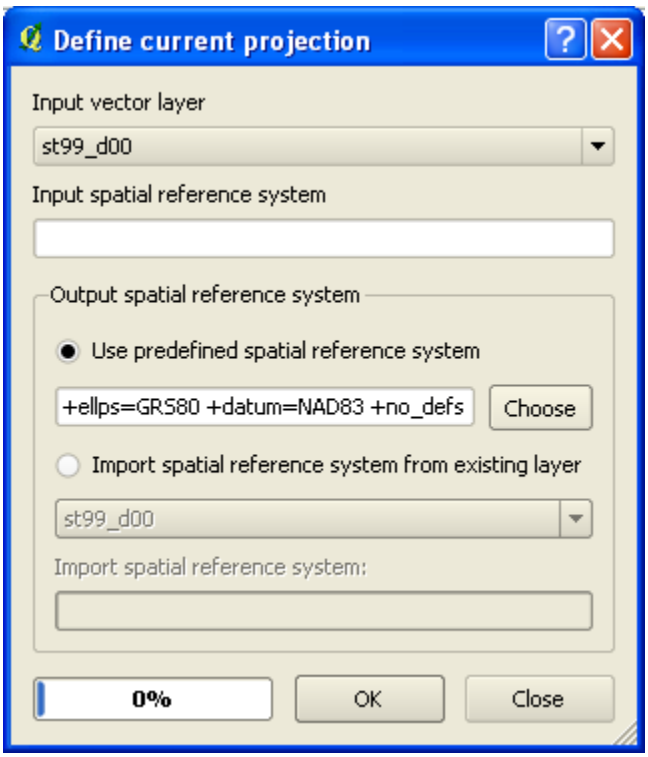
This section will show you how to define a projection for data that is missing this information, and then how to transform a file from one projection to another. Choosing a coordinate system and map projection for your layers is of critical importance; all layers in a project need to share the same system in order to work together, and the choice of a system is influenced by the type of analysis you're doing and what your final map will depict.

Steps

1. *Create a new project.* Open QGIS to an empty, blank project. Hit the  Save As button. Browse to your data folder for part 4 and save the project as part4.qgs. We'll be working with this project throughout this part of the tutorial.
2. *Define the project window.* On the menu bar go to Settings > Project Properties, scroll through the coordinate system list and select NAD 83 as the coordinate reference system (CRS). Click OK.
3. *Check the shapefile's CRS.* Minimize QGIS, and use your file browser to browse through the data folder for part 4. You'll see there's a shapefile in the folder called st99_d00.shp, and it has a .dbf and .shx file associated with it. This file represents the states of the United States. However, the projection file, .prj, is missing. Normally we could open the .prj file in a text editor and see what the projection is. In this case the file is undefined; occasionally you'll download files where the prj is missing. We have to go back to the source where we downloaded the file and check the metadata to see how it's defined (it's from the US Census Bureau, and according to the site the files were created in NAD 83), and then we must explicitly define it in QGIS.
4. *Add the states shapefile.* Maximize QGIS. Hit the  Add Vector data button. Browse to the part4 data folder and add the st99_d00 shapefile. Use the  Zoom In button, draw a box around the US and zoom in.



5. *Define the projection.* On the menu bar go to Vector > Data Management Tools > Define current projection. Keep st99_d00 as the input vector layer. In the Output spatial reference system area hit the Choose button, expand the Geographic Coordinate System menu, scroll through the CRS list, select NAD 83 and hit OK. Then on the Define current projection menu hit OK. Close the menu when you're finished.



6. *Check the prj file.* Minimize QGIS and go back to your data folder for part 4 using your file manager. You should now see a file called st99_d00.prj. Open this file in a text editor (if using Windows, select the file, right click, choose the option to select a program from the list, select Notepad and click OK). You will see the projection information stored in the file:

```
GEOGCS["NAD83",
    DATUM["North_American_Datum_1983",
        SPHEROID["GRS 1980",6378137,298.257222101,
            AUTHORITY["EPSG","7019"]],
        AUTHORITY["EPSG","6269"]],
    PRIMEM["Greenwich",0,
        AUTHORITY["EPSG","8901"]],
```

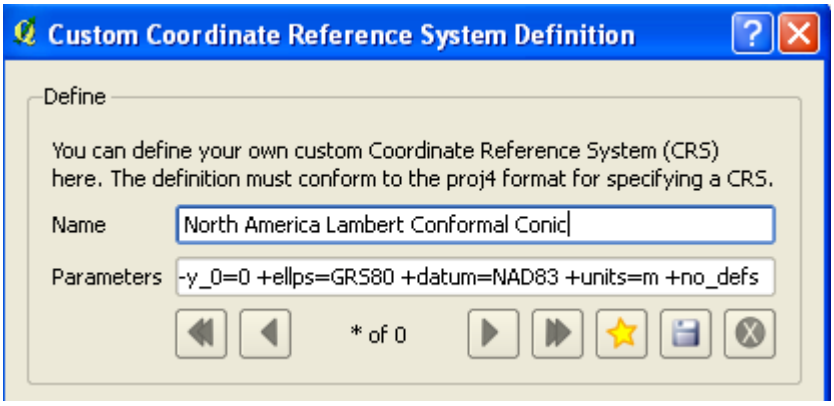
```
UNIT["degree",0.01745329251994328,
    AUTHORITY["EPSG","9122"]],
AUTHORITY["EPSG","4269"]]
```

. Close the file when you're finished.

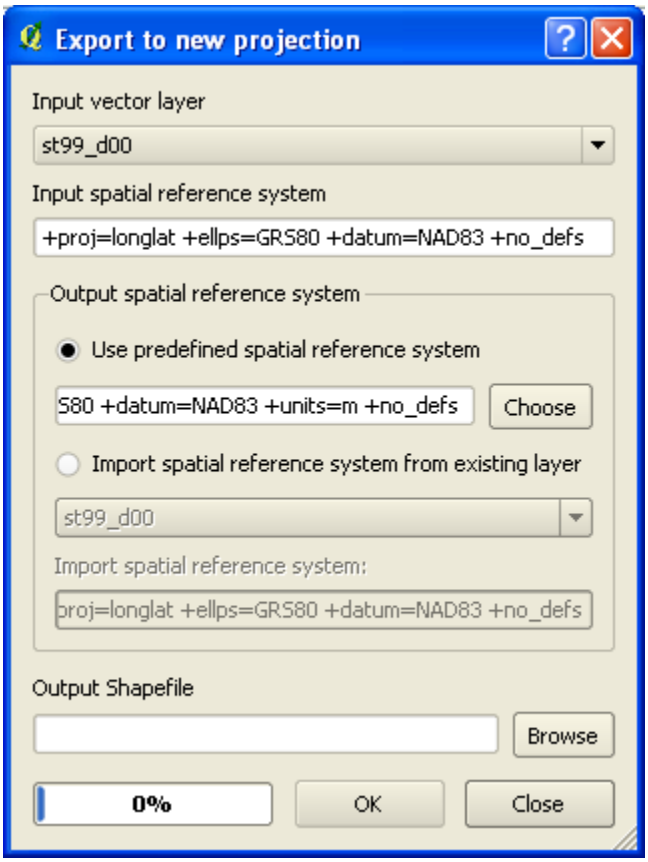
7. *Define a custom projection.* Now that the layer has been properly defined based on what it currently is, we can now transform the layer to a different system that's more suitable for a thematic map of the US. Instead of using NAD83, which is a basic geographic coordinate system (GCS), we are going to use a projected coordinate system (PCS). While QGIS has access to a large number of GCS's in its library, it doesn't not have many PCS's that are suitable for continental or global projections. We need to add a custom projection to the QGIS library. In the data folder for part 4, open the file called lcc_na_proj4.txt. Copy the contents of the file:

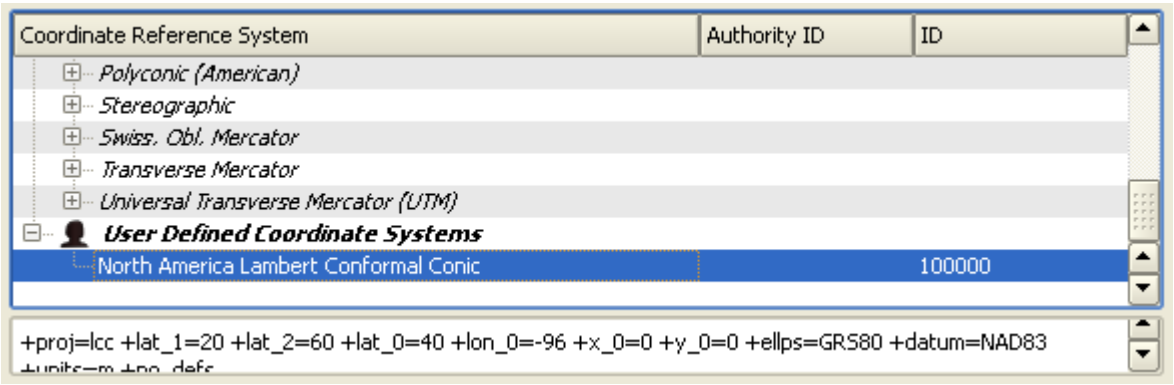
   +proj=lcc +lat_1=20 +lat_2=60 +lat_0=40 +lon_0=-96 +x_0=0 +y_0=0 +ellps=GRS80 +datum=NAD83 +units=m +no_defs.

   Maximize QGIS. On the menu bar go to Settings > Custom CRS. Paste the projection information from the file into the Parameters box. In the Name box, type North America Lambert Conformal Conic. Hit the Save Button, then hit OK to close the window.

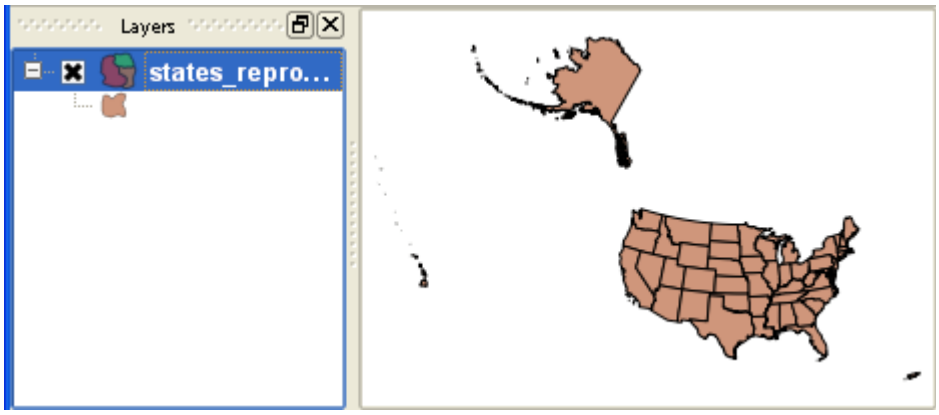


8. *Transform the projection.* On the menu bar go to Vector > Data Management Tools > Export to new projection. Keep the input layer as st99_d00. Under output spatial reference stsyem, hit Choose. Scroll down to the bottom of the CRS and open the menu that says User Defined Coordinate Systems. Select the North America Lambert Conformal Conic projection and hit OK. For the output shapefile, browse to your data folder for part 4 and name the file states_reproject.shp Hit OK. Click yes to add the new file to the map. Close the export window.





9. *Change the definition for the window.* At this point you'll notice something strange; your layers will not be able to draw properly because you now have two layers that don't share the same CRS. Select the old st99_d00 layer in the map legend (ML), right click and remove it. Select the state_reproject layer in the ML and hit the  Zoom to Layer button. You should see your newly projected layer. On the menu bar go to Settings > Project Properties, scroll through the coordinate system list and select the North American Lambert Conformal Conic CRS in the Custom projection section. Click OK. Hit the  Save button.



Commentary

Understanding Coordinate Reference Systems

All GIS layers are created using a specific coordinate reference system (CRS). The reason that we can take data from different sources and overlay them in GIS is because they share the same system; likewise, we can plot coordinate data and create layers because there's a coordinate system under the hood of our map window. In order for everything to work, your layers must share the same system and the map window must be defined to use that system. GIS software can be used to transform layers from one system to another. Each CRS is composed of at least three or four parts:

Spheroid or Ellipsoid: We typically imagine the earth as a perfectly round sphere, but in reality the earth is rather lumpy and uneven, with protrusions in some areas and indentations in others. The shape of the earth is approximated using spheroids, round three dimensional models of the earth, and ellipsoids, which represent the earth as being more oval than sphere-like in nature.

Coordinate System: This is the reference grid used for locating places on the earth and measuring distances. Latitude and longitude is the most common system, but there are other systems with different grid cells and units of measure; for example, the Universal Transverse Mercator (UTM) system uses a unique grid.

Datum: When you apply a coordinate system like latitude and longitude to different spheroids or ellipsoids, there needs to be a method for creating the grid and attaching it to the earth's surface. Mathematically, where does one draw the prime meridian and equator on a particular spheroid in order to accurately represent their location? The frame of reference for drawing these lines and measuring locations on the surface of the earth is called a datum.

Collectively, when you have these three elements: a spheroid or ellipsoid, a datum, and a coordinate system, you have something called a Geographic Coordinate System (GCS). The terminology is confusing, as a coordinate system is one part of a geographic coordinate system, and most systems are named based on the datum they use. For example, WGS84 (World Geodetic System of 1984) is the most common GCS and uses the GRS 1980 spheroid, WGS84 as a datum, and latitude and longitude as a coordinate system. WGS84 is used by the Global Positioning System of satellites and thus by individual GPS units as a default, and is commonly used by online mapping applications. It is so common that it is often referred to a THE Geographic Coordinate System. There are other systems; in North America NAD83 (North American Datum of 1983) is widely used, particularly by government agencies.

If you add a map projection as the fourth element to the spheroid/ellipsoid, datum, coordinate system trio, you have a projected coordinate system (PCS):

Projection: Map Projections are mathematical systems for taking the three dimensional earth and transforming it to a flat two dimensional surface. There is no way to take a 3D shape and accurately represent it on a 2D surface, so map projections are designed to preserve one quality of the earth - area, shape, or

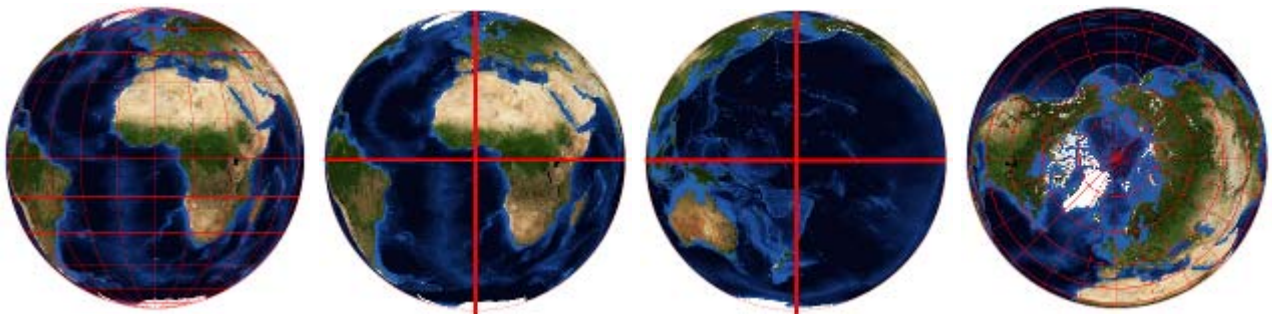
distance/direction, or are created as a compromise to make the earth appear the way we expect it to appear on a flat surface.

It's important to understand the distinction between a GCS and a PCS, because when you go to transform a layer or define a projection these two systems will be stored or organized in the software separately, under different menus or tabs. You should use a GCS when you're doing analysis, measuring distances, or working in a relatively small geographic area. You should use a PCS when you're creating a thematic map or need to have a certain quality of the earth (area or shape) preserved.

Latitude and Longitude

The most common coordinate system is latitude and longitude, a grid system that covers the earth and uses a unit of measurement called a degree. Lines of latitude, called parallels, run east-west. The origin of latitude is the equator, which is zero degrees latitude. The equator bisects the earth and along this line there are twelve hours of daylight and twelve hours of darkness each day, throughout the year. Lines of latitude run 90 degrees to the north pole and 90 degrees to the south pole. One degree of latitude is equal to approximately sixty-nine miles, and since they are parallel lines they never converge.

Lines of longitude, called meridians, run north-south. Unlike the equator, which is the defacto line of latitude based on natural phenomena, the selection of an origin for longitude is arbitrary. The Prime Meridian, zero degrees longitude, was designated as the origin parallel in the 19th century. It runs through the center of the astronomical observatory in Greenwich, UK. There are 180 degrees of longitude to the east and to the west of the prime meridian. The meridian that is opposite the prime meridian on the far side of the globe, 180 degrees longitude, is the International Date Line. Unlike latitude, longitude converges at the poles to a single point at zero degrees. Since lines of longitude converge there isn't a uniform distance between them - the distance decreases as you move away from the equator. At the equator one degree of longitude is approximately 69 miles across, while at the poles it is zero miles.



There are two conventions for recording coordinates: in degrees, minutes, and seconds (DMS) or as decimal degrees (DEC). Take a look at the following coordinates for Philadelphia, PA from the USGS GNIS gazetteer:

39 deg 57' 08" N 75 deg 9' 50" W (DMS)

39.952335, -75.163789 (DEC)

The DMS notation is similar to the notation for telling time - there are 60 minutes in one degree and 60 seconds in one minute. DEC notation is preferable for computer processing; if you're plotting coordinates in GIS they should be in DEC. In DEC, latitudes south of the equator and longitude west from the prime meridian to the international date line are recorded as negative numbers. It is crucial that DEC coordinates indicate direction, otherwise you'll be confusing your point with a different place:

39.952335, -75.163789 is Philadelphia, PA USA

39.952335, 75.163789 is a remote area in western China near the Kyrgyzstan border

Map Projections

Most people today would agree that the earth is round. Most maps, whether they're on paper or a computer screen, are flat. When you take a three dimensional sphere and flatten it to two dimensions, you get fair amount of distortion. Imagine removing the peel from an orange and laying it out flat - you can't do it without tearing the peel. A map projection is a method for taking the three dimensional earth and transforming it to a flat surface.

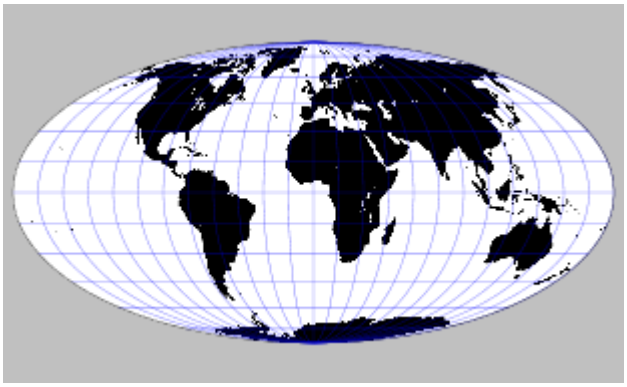
For a nice overview, visit Radical Cartography's projection page and note the common projections (marked in pink). Projections can be classified based on how the grid is applied to the earth's surface - a grid laid flat on top (azimuthal), wrapped as a cone on the top half of the earth (conical), wrapped around the earth as a cylinder (cylindrical), etc. They can also be organized based on which property they preserve:

- Area - areas that are the same size on the globe appear as the same size on a map. Examples: Mollweide projection for the earth, Albers Equal Area for continents
- Shape - preserves angles around points on a map, and therefore preserves shapes for small to medium areas. Examples: Mercator for the world, Lambert Conformal for continents
- Distance / Direction - a straight line on the map will give you the shortest distance between two points, the same distance as a great circle on a globe. The Geographic Projection, also known as Geographic Coordinate System (GCS) or Plate Carree, is the most common

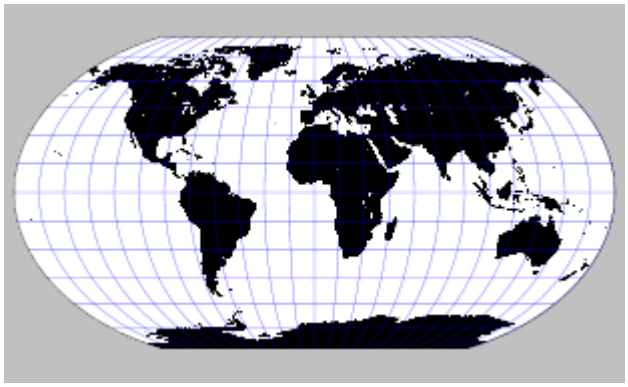
Other projections:

- Interruptions - these projections show tears in the earth's surface and try to mitigate them to create something readable. Goode's Homolosine is good for showing land areas, but poor for showing oceans (as these are interrupted).
- Compromises - these projections don't preserve any quality of the earth exactly, but they compromise to make a map of the earth that "looks right". Good compromise projections of the earth include Robinson and Winkel Tripel.

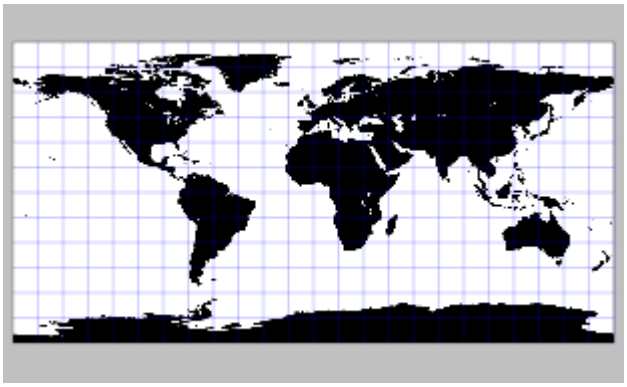
You can compare maps that use different projections to get a sense for how they distort different areas (in particular, observe Greenland):



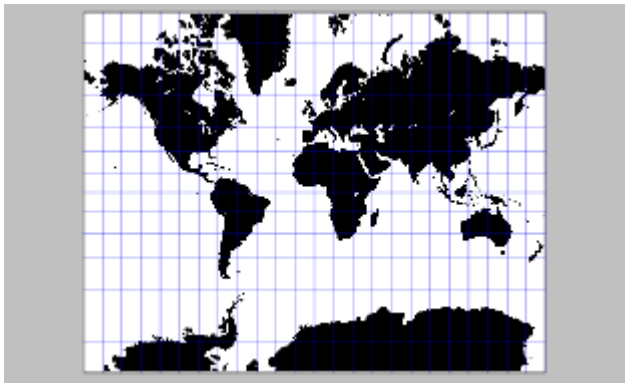
Mollweide



Robinson



GCS (aka Plate Caree)



Mercator

Common map projections for the world for general reference or thematic use include Robinson, Mollweide, Goode Homolosine, and Winkel Tripel (the first two have proj4 definitions that can be custom defined in QGIS). In general, projections that appear oval-like, showing the curvature of the earth at the edges, are best for general or thematic use.

Every continent and country has a preferred map projection or set of projections that is appropriate for each area based on its size and shape. Look at atlases or pre-existing maps to get an idea of what these are. Albers Equal Area, Lambert Equal Area, and Lambert Conformal are common and are adjusted to focus on specific continents or countries. Orthographic projections are used to map polar areas.

GCS Definitions

Several formats have been created for recording the definition of projections. There's the Open Geospatial Consortium's Well-Known Text Format (OGC WKT) as seen in the example we worked through, the Proj4 format, which we used to define a custom CRS in QGIS, and .prj file format created by ESRI. To look up CRS information, you can use the spatial reference website at http://spatialreference.org/. Use that site to get the proj4 format for creating custom projections in QGIS.

Geographic Reference Systems have also been classified with codes, which makes them easier to identify and retrieve. The QGIS CRS draws it's systems from a library called the European Petroleum Services Group (EPSG). This library contains most of the primary GCS systems, such as WGS84 and NAD83, and local PCS systems like State Plane. For example, EPSG 4269 is the code for NAD 83, and EPSG 4326 is the code for WGS 84. The advantage of the codes is clearer when you're working with longer names: NAD 83 NY State Plane Long Island is abbreviated to EPSG 32118. The EPSG library lacks most of the PCS systems for continental and global map projections, which is why these are not available in QGIS; search Spatial Reference to find the proj4 definitions for these projections in order to custom define them.

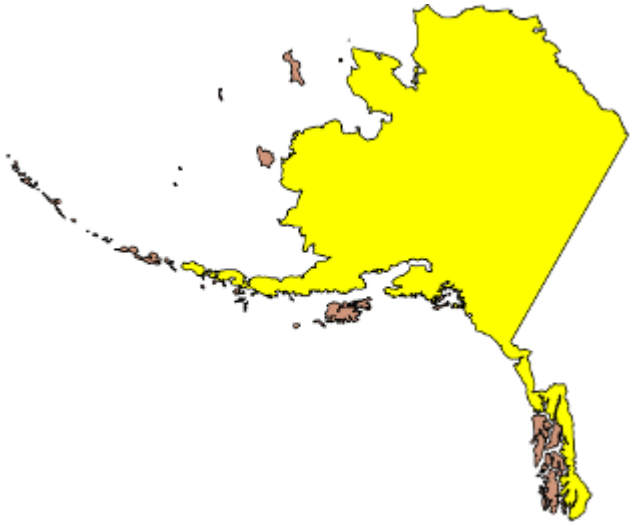
Section II: More Geoprocessing

This section will demonstrate a few more geoprocessing techniques that you're likely to need. You'll learn how to convert a single part layer to a multi-part layer and will do another table join.

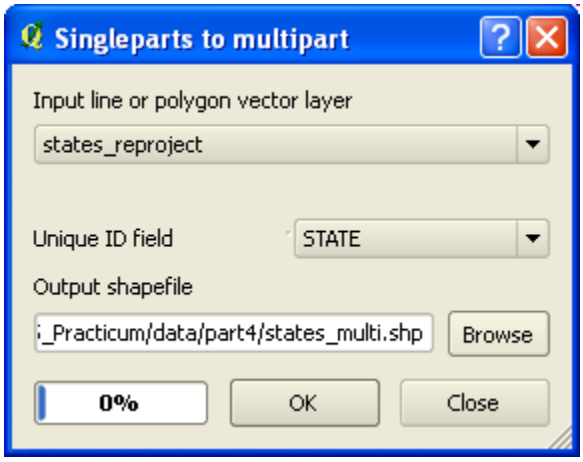
Steps

1. *Examine the attribute table.* Select the states_reproject layer in the ML, right click and open the attribute table. In the top left-hand corner of the window, it says that the attribute table states_reproject has 273 features. If there are 50 states, plus two (DC and Puerto Rico), how could there be 273 records?
2. *Examine a selection.* Notice that there are several records for Alaska. Click on the first record for Alaska in the table

to select it. Close the attribute table. Pan your map view to see Alaska. Notice that one large portion of the state is selected, but none of the islands that are separate from the mainland are. This shapefile is a single-part file, meaning that each individual polygon is an independent feature with its own record in the attribute table. You can select other islands in Alaska with the  Select feature tool to test this. When you're finished,  clear all selected features.



3. *Convert to multi-part feature.* Before we join an attribute table to our shapefile, we need to convert the layer to a multi-part feature - a layer where a single feature (state) can be made up of multiple polygons. This will allow us to do a one to one join between the state layer and the attribute table. On the menu bar go to Vector > Geometry Tools > Singleparts to mulitpart. The input file will be states_reproject. The Unique ID field that will be used to convert the file (associating individual polygons with a feature) is the STATE field. Browse to the data folder for part4 and save the file as states_multi.shp. Click OK. Click Yes to add the new layer to the project. Close the single to multi menu.



4. *Open the new attribute table.* Open the attribute table for the new states_multi layer. You'll see that there are now 52 records in the layer, which is what we're expecting. Close the attribute table. Select the old states_reproject layer in the ML, right click and remove it. Hit the  Save button.
5. *Examine the employment attribute table.* Minimize QGIS. Use your file browser to go to the part 4 data folder. Find the file called info_emp_bls_2008.dbf. Right click on the file and open it with a a spreadsheet program. There are 51 records, one for each state and DC. The CODE field is a FIPS code we can use for joining, EMP51 is the number of people who are employed in the Information Sector, as defined by the North American Industrial Classification System, and TOTAL_EMP is the total number of people in the labor force. Close the file. Maximize QGIS. (Note: if

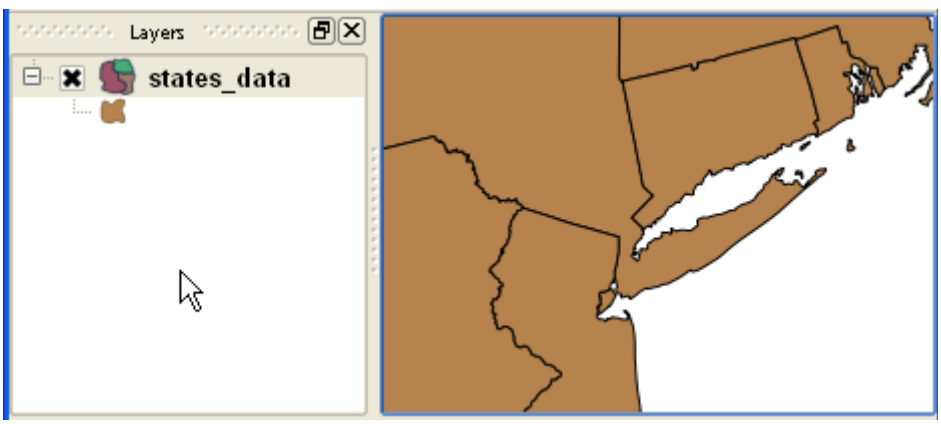
you're using QGIS version 1.6, you could add the data table to QGIS to view it by adding it as a vector layer).

	A	B	C	D	E	F
1	CODE	STATE	ID_NAICS51	NAICS51	ID_TOTAL	TOTAL_EMP
2	01	AL	ENU0100010551	27013	ENU0100010510	1570188
3	02	AK	ENU0200010551	6988	ENU0200010510	237708
4	04	AZ	ENU0400010551	41833	ENU0400010510	2174919
5	05	AR	ENU0500010551	18653	ENU0500010510	971640
6	06	CA	ENU0600010551	467864	ENU0600010510	13039293
7	08	CO	ENU0800010551	76935	ENU0800010510	1943153

info_emp_bls_2008

6. *Join the data to the shapefile.* On the menu bar go to Vector > Data Management Tools > Join Attributes. states_multi is the target vector layer. The target join field in this layer is STATE (contains the two digit FIPS code). In the Join data section select Join DBF table. Browse to the part 4 data folder and select info_emp_bls_2008.dbf. Specify CODE for the Join field. Browse and save the output shapefile to the part 4 data folder as states_data. Under output table, keep the default that says Only keep matching records. Hit OK. Click Yes and add the new file to the map. Close the menu.

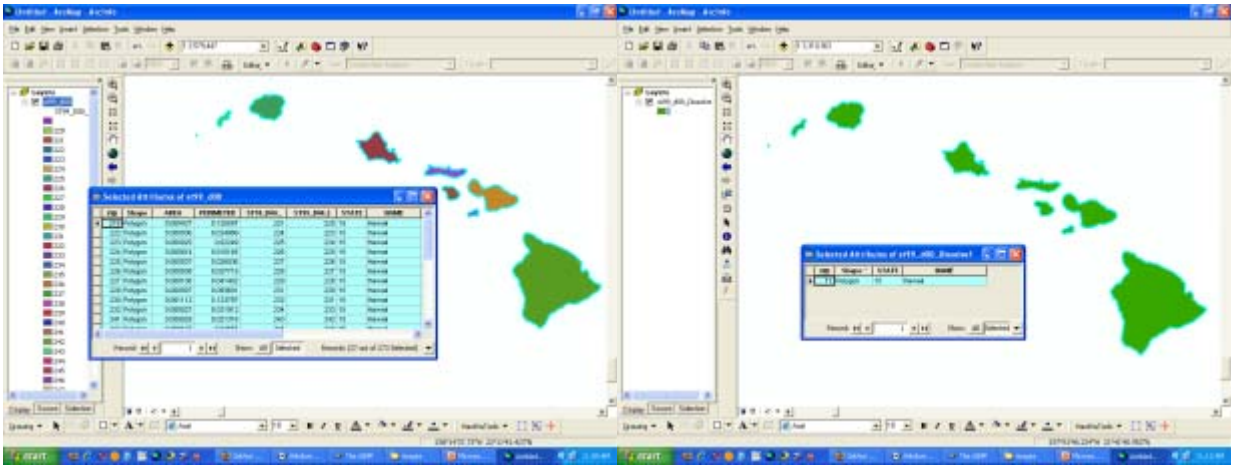
7. *Remove the old layer and inspect the new one.* Select the states_multi layer in the ML, right click and remove it. Notice that the states_data layer no longer contains Puerto Rico; since there was no data in the employment attribute table for Puerto Rico the feature was dropped from the shapefile after the join. Select the states_data layer in the ML and open the attribute table. You'll see there are only 51 records, and the data from the employment table has been appended to the file. Close the attribute table. Save your project.
8. *Inspect the new layer.* Zoom in to the northeastern US, to the area around New York City. You'll notice that, unlike the previous census file we worked with from TIGER, this file has already been modified to remove bodies of water from state boundaries. But if you look at the NYC area, you'll see that Manhattan and Long Island appear joined to the mainland. This shapefile is from the Census Generalized Cartographic Boundary Files; they are TIGER files that have had their boundaries simplified so they appear less jagged at small scales (viewing the US as a whole) but are not appropriate for large scale maps (viewing a small area like the NYC metro).



Commentary

Singlepart and Multipart Features

Polygon features in shapefiles or other vector formats may consist of multi-part or single part features. With single-part features, each individual polygon has a record in the attribute table. With multi-part features, each feature has a record in the attribute table regardless of how many polygons make up the feature. For example, in a single-part shapefile of Hawaii each island has it's own record in the attribute table (1st image below), whereas in a multi-part shapefile all of the islands are combined into a single feature, the State of Hawaii, for which there is one record in the attribute table (2nd image below). Most GIS systems have tools for converting one format to another - this is important, because if you want to join a data table to a shapefile, you'll usually want it to be a multi-part shapefile. For example, if you are joining a state-based data table to a shapefile, it wouldn't make sense to assign the entire population of Hawaii to each individual island - it would lead to errors when classifying data and calculating statistics.



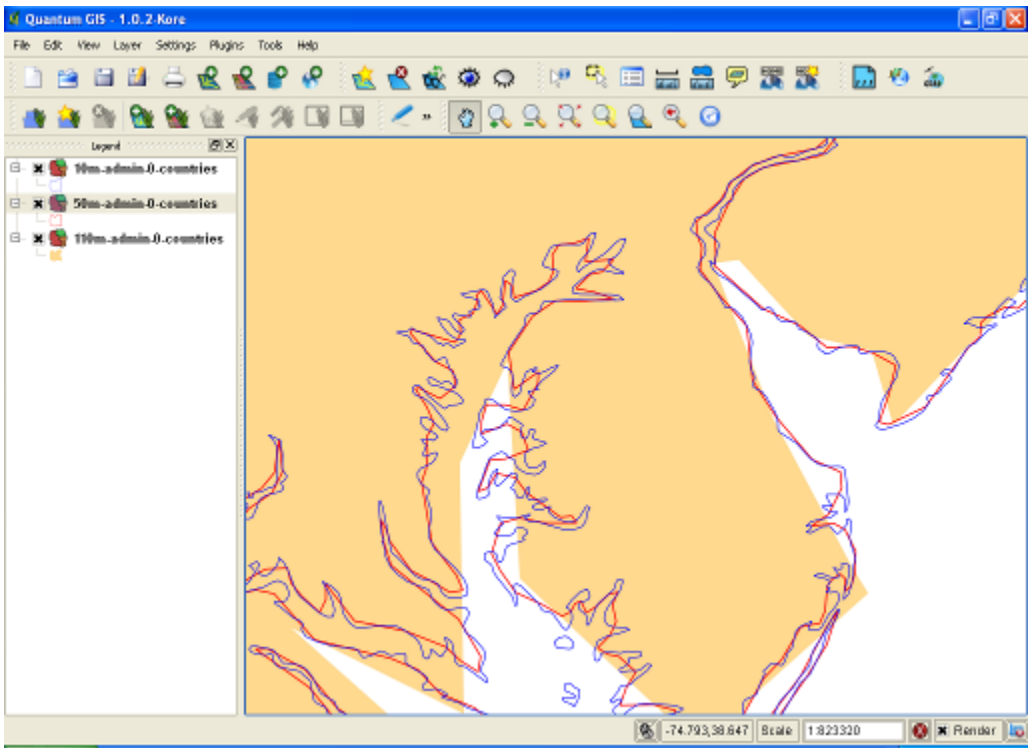
Generalization and Scale

The Census Generalized Cartographic Boundary Files http://www.census.gov/geo/www/cob/bdy_files.html that we are using in this part of the tutorial were designed for creating maps of the US at a national or regional scale. According to the Census Bureau, "The cartographic boundary files are primarily designed for small scale, thematic mapping applications at a target scale range of 1:500,000 to 1:5,000,000." Boundaries have been generalized to depict land areas, to smooth coastlines and boundaries, and to remove small islands. This makes the boundaries appear smoother and cleaner at these smaller scales, while sacrificing accuracy that wouldn't be visible.

When choosing vector files for thematic mapping you will need to make sure that the generalization for the file is appropriate for the scale you're working at. If you were creating a map of the NYC metro area, you would not want to use these boundary files as the generalizations become apparent at this larger scale and will make your maps appear inaccurate. You can identify whether a layer is appropriate by looking at the metadata and seeing if an optimal scale is indicated. Scale is a proportion of units of measurement on the map versus the actual distance in reality. A scale of 1:5,000,000 indicates that one measurement unit on the map represents 5,000,000 units in reality. Small scale maps cover large areas while large scale maps cover small areas; this may seem counter-intuitive, but remember that scales represent fractions: 1/5,000 is a larger number (and thus larger scale) than 1/5,000,000. Most GIS software have tools for generalizing boundaries if you need them to be more simplified.

The screenshot below illustrates differences in generalization in scale using vector data from the Natural Earth website, which provides free generalized vector data for creating professional medium to small scale maps. The map below is of the Delmarva peninsula and shows an overlay of 3 different shapefiles created at 3

different scales. Scales range from small to large and from most to least generalized: the beige area is a 1:110 mil scale, the red line is 1:50 mil scale, and the blue line is 1:10 mil scale. Obviously you wouldn't want to use the 1:110 scale layer depicted in beige to create a map of this area as it is far too generalized, but it would be well-suited for a national map.

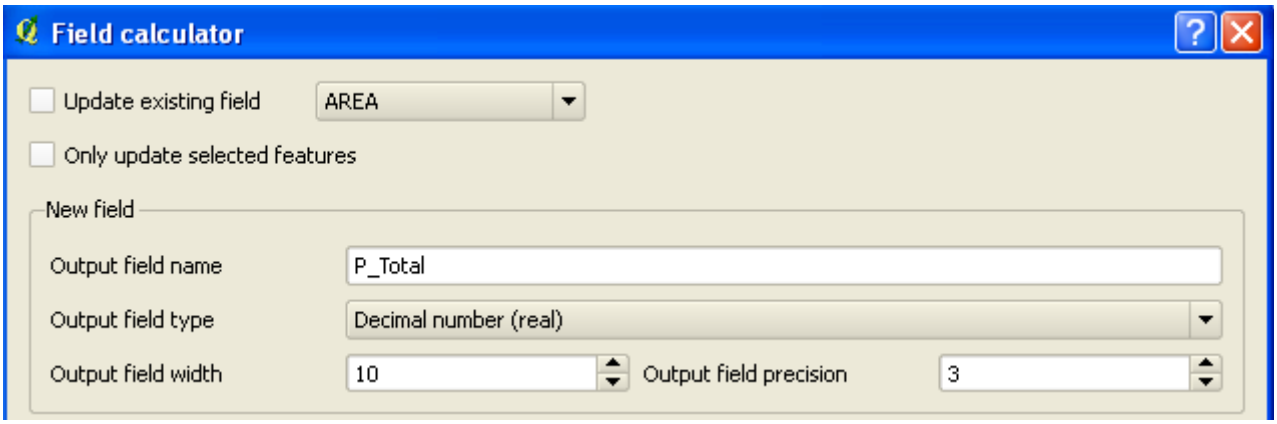


Section III: Creating Calculated Fields

This section will show you how to add new calculated fields to a shapefile in QGIS. In many instances mapping numbers that represent whole values may not make sense and you'll want to create derived values. In this section you'll create a percent total and a location quotient to show the concentration of employment in the information sector across different states.

Steps

1. *Enter the edit mode.* Select the states_data layer in the ML, right click and open the attribute table. Hit the  Edit button below the table to enter the edit mode. Since we are making changes to the actual shapefile we need to do that from an edit mode.
2. *Launch the field calculator.* Hit the field calculator button that's a few buttons to the right of the edit button. This opens the Field Calculator window. Under New Field, type P_Total as the output field. Change the field type to a Decimal number (real). Keep the output field width to 10 (default width setting in the table window) but change the precision to 3 (number of places right of the decimal point). In the fields box, click NAICS51 to add it to the expression field. Hit the divisor symbol under the operators. Then click TOTAL_EMP in the Fields box. Your field expression should read NAICS51 / TOTAL_EMP. Hit OK. Back on the attribute table screen, hit the  Edit button to stop editing and save your edits. You'll see the new percent total field appended on the right.



3. *Create another calculated field for location quotient.* Enter an edit mode and repeat the previous step to create another calculated field. Name the field Locat_Q for location quotient. Use the same field width and precision as before. You'll create the expression by clicking on field names, operators, and typing numbers directly into the field box. Your final expression should look like this: (NAICS51 / 2989162) / (TOTAL_EMP / 113188646). Be sure to include the parentheses and omit commas. The first number represents total employment in the information sector for the country, and the second number represents the country's total labor force (you could calculate these numbers using the Basic Statistics tool under Vector > Analysis Tools). Save your edits when finished.

STATE_2	ID_NAICS51	NAICS51	ID_TOTAL	TOTAL_EMP	P_Total	Locat_Q
AK	ENU0200010551	6988	ENU0200010510	237708	0.029	1.113
MN	ENU2700010551	57553	ENU2700010510	2304189	0.025	0.946
WA	ENU5300010551	104932	ENU5300010510	2429793	0.043	1.635
MT	ENU3000010551	7624	ENU3000010510	356638	0.021	0.809
ID	ENU1600010551	11664	ENU1600010510	539153	0.022	0.819
ND	ENU3800010551	7430	ENU3800010510	286096	0.026	0.983

4. *Examine your data.* Click on the Locat_Q column to sort the data. States with a location quotient higher than 1.0 have a higher concentration of employment in the Information Industry relative to other parts of the country, while states with quotients lower than 1.0 have a smaller concentration relative to the rest of the country. When you're finished examining the data close the table. You can  Save your project at this point, but the edits to your shapefile have already been saved, since we were working on the shapefile directly and saved it after the edit mode.

Commentary

Representing Values

In some circumstances it may make sense to map values as whole numbers - cities by number of crimes, states by total population, counties by number of renter-occupied housing units, etc. But in each of these examples a particular place could have a higher value simply because it has more people or is a larger place. In order to make more meaningful comparisons it's often necessary to do a little math:

- Percentage - (value of subset / total value)*100: (3,000 renter units / 10,000 renter units)*100 = 30% units are rentals
- Rate - (value / total value) * multiplier: (400 robberies / 50,000 people)*100,000 people = 800 robberies per 100,000 people
- Ratio - (value 1 / value 2): (4000 cars / 3000 people) = 1.33 cars per person

- Density - (value / land area): (800,000 people / 2500 sq miles) = 320 people per sq mile
- Percent Change - [(recent value / older value)-1] * 100: [(10,000 people / 9,000 people)-1] * 100 = 11.1% change

A location quotient (also known as a location coefficient) is a common method used in economic base analysis. It compares a local economy to the greater economy in order to measure how specialized a local economy is for particular industries. To calculate a location quotient:

(employment in industry in local economy / total employment in local economy) / (employment in industry in national economy / total employment in national economy)

- A value > 1.0 indicates that the local economy is more specialized in a particular industry relative to the nation
- A value < 1.0 indicates that the local economy is less specialized in a particular industry relative to the nation
- A value = to 1.0 indicates that the local economy for that particular industry is average relative to the nation

The data used in our example is from the US Bureau of Labor Statistics at http://www.bls.gov/.

Industrial Classification: NAICS

The North American Industrial Classification System (NAICS) is a hierarchial system of codes used to classify businesses into industries in the US, Canada, and Mexico. It was created in the mid 1990s and replaced the older Standard Industrial Classification (SIC) system. The NAICS system consists of broad industrial sectors defined with two digits that can be broken down into more specific subsectors with additional digits.

In our example we are studying the labor force of NAICS 51, the Information Sector. Establishments in NAICS 51 produce and distribute information and cultural products, provide the means to transmit or distribute these products as well as data or communications, and process data. NAICS 51 can be broken down into more specific parts that include:

- 511 Publishing Industries except the Internet
- 512 Motion Pictures and Sound Recording
- 515 Broadcasting except the Internet
- 517 Telecommunications
- 518 Data Processing, Hosting, and Related Services
- 519 Other Information Services

Each of these 3 digit subsectors can be broken down into 4 digit groups (515 Broadcasting can be broken down to 5151 Radio and TV broadcasting and 5152 Cable and Other Subscription Programming), 4 digit groups can be broken down to 5 digit industries (5151 breaks down to 51511 Radio Broadcasting and 51512 TV Broadcasting), and 5 digit industries can be broken down to 6 digit national industries (51511 breaks down to 515111 Radio Networks and 515112 Radio Stations).

You can browse and download the codes at http://www.census.gov/eos/www/naics/. They are widely used by government agencies that produce data for industries (US Bureau of Labor Statistics, US Census Bureau, Statistics Canada, National Institute of Statistics Mexico) as well as private companies that produce databases or information retrieval systems that focus on industrial research. The NAICS system is largely compatible with the UN Statistics Division's International Standard Industrial Classification (ISIC) codes.

Section IV: Classifying and Symbolizing Data

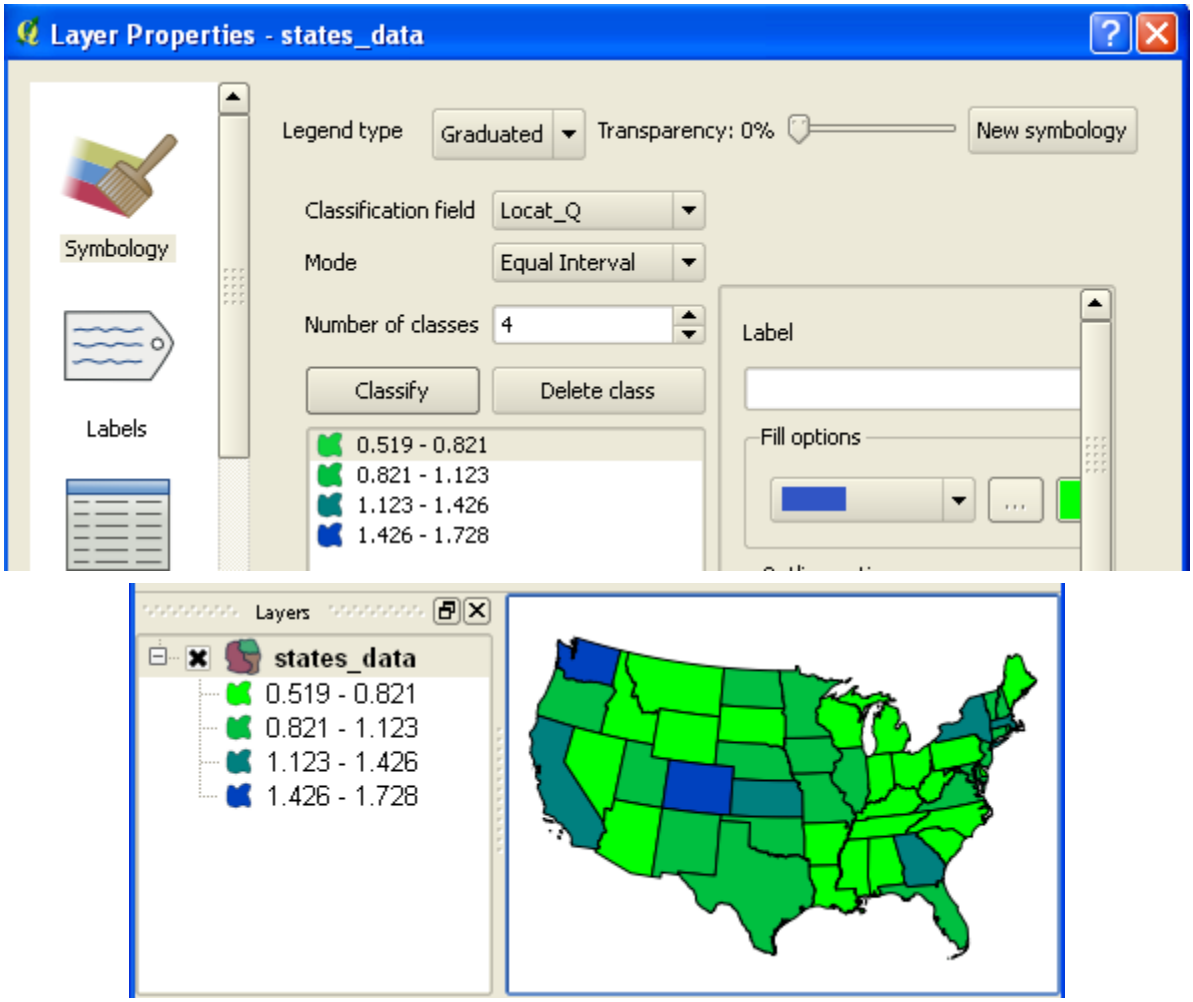
In this section you'll learn about the different methods for classifying data and the best approach for

choosing color schemes to symbolize your data. These are important concepts to grasp, as they have a direct impact on how successful your map will be in communicating your data.

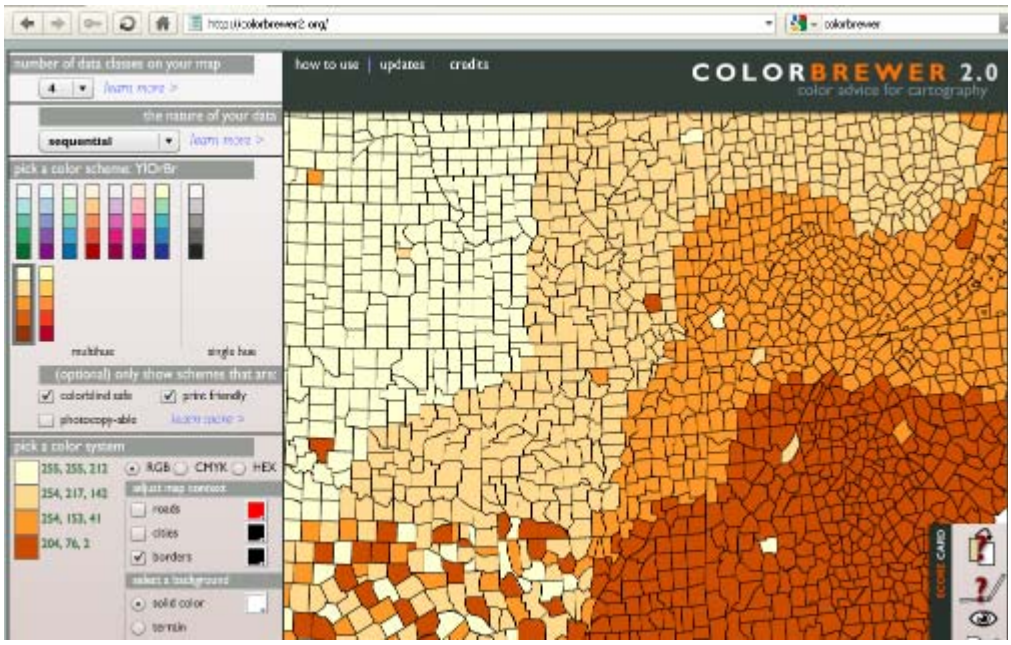
In this section we'll work with the "Old symbology" tab that has the standard functionality QGIS has had for quite some time. In the commentary for this section you can read about the "New symbology" tab as of version 1.6 - the old tab is still the default in 1.6 but the new tab is also available, and can make certain aspects of the symbolization process easier.

Steps

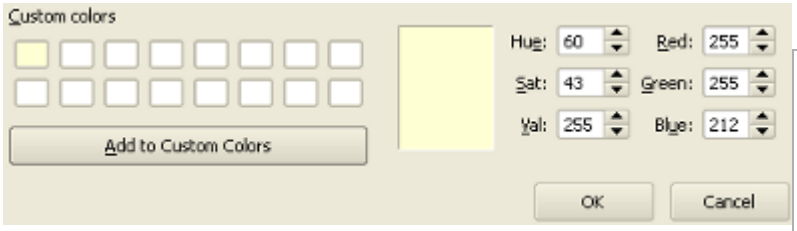
1. *Classify your data*. Select states_data in the ML and double-click to open the properties menu. Go to the Symbology tab. Change the legend type dropdown to Graduated Symbol. Change the classification field dropdown to Locat_Q. Keep the mode set to Equal Intervals. Change the number of classes to 4. Hit the Classify button. Hit OK, and take a look at the result in the map window.



2. *Visit Colorbrewer to find a good color scheme.* The default color scheme for QGIS is not very good, and you shouldn't use it to produce final maps. Use the Colorbrewer tool at http://colorbrewer2.org/ to pick a good scheme. Choose 4 classes for your map. Keep the option for a sequential color scheme. Check the boxes to limit your options to schemes that are color blind safe and print friendly. Click on different color bar options until you find one you like. In the lower-right hand corner of the map, you can click on a scorecard that shows whether your choice is ideal for the color blind, color printing, photocopying, and viewing on an LCD screen.



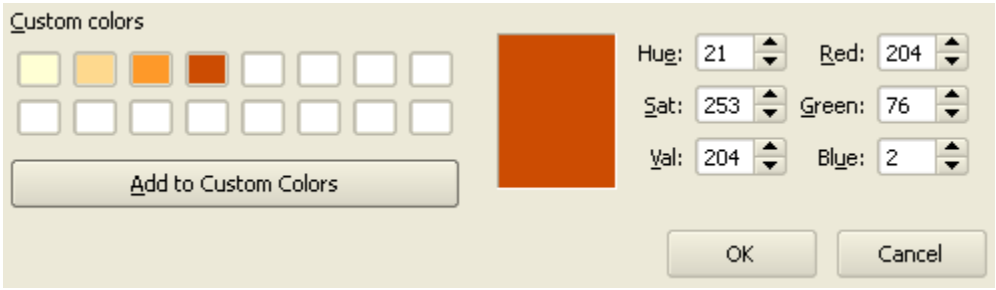
3. *Add custom color to QGIS* In QGIS, return to the Symbology menu for your states_data. Select your first category of data in the classification range. Then, in the adjacent box that contains the Fill Options, click on the current color (it should be bright green). This opens the Select Color menu. Click on the first empty box under Add Custom Colors. Then, in the dropdowns on the right that have values for Red, Green, and Blue, change each of these values to match the lightest color in the color scheme you chose on Colorbrewer that are listed under pick a color (you can flip back and forth between the color window and Colorbrewer, or jot down all the color codes from Colorbrewer on a piece of paper, or use the sample codes below). Once you have entered the three numbers, click Add to Custom Colors to add it to the first box.



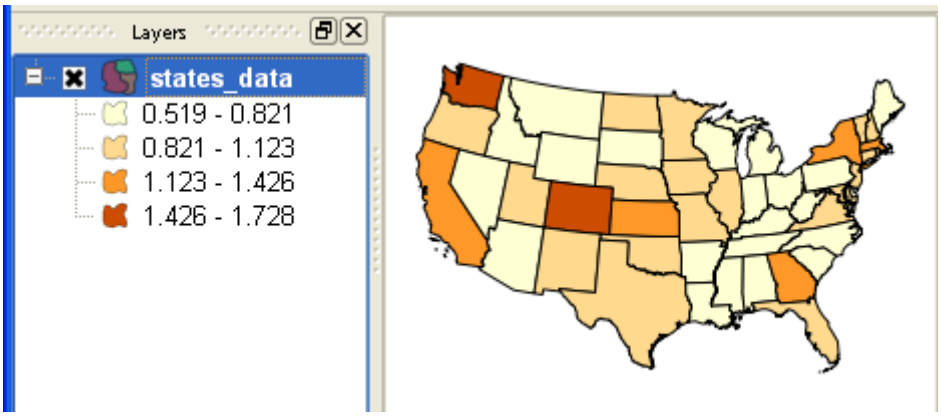
4 Class Color Schemes

Colors	RGB Values (Light to Dark)			
Oranges	255 255 212	254 217 142	254 153 41	204 76 2
Greens	237 248 251	178 226 226	102 194 194	35 139 69
Pink to Purple	254 235 226	251 180 185	267 104 161	174 1 126

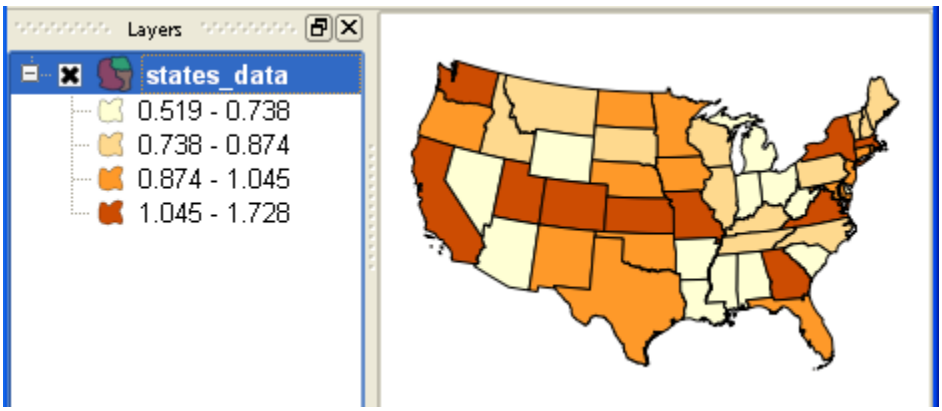
4. *Add the rest of your colors to custom colors.* Select the next empty box under the custom colors box (it's important you do this first, otherwise you'll overwrite the color in the first box). Repeat the step above to add all three colors to your custom menu. When you're finished, click OK.



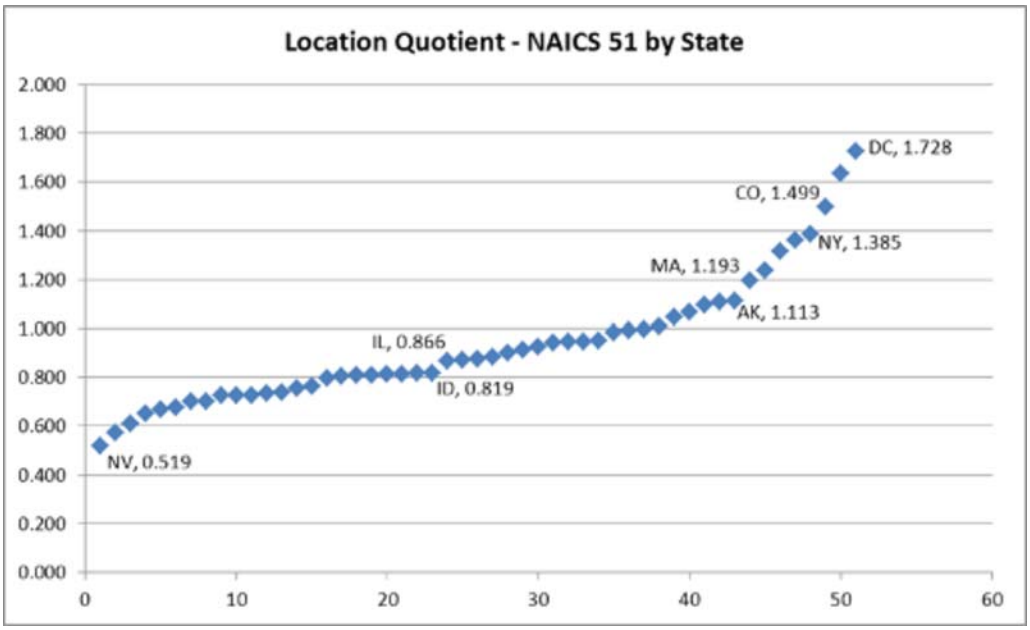
5. *Assign custom color to your data.* Select the first data range in your classification. Click on the fill color. Change the color to the first custom color in your menu (the lightest shade) by clicking on its color box and click OK. Select the second data range in your classification. Click on its fill color, and change it to the second custom color in your range. Click OK. Repeat these steps for your other colors, so each range is assigned to its appropriate color. When you're finished, click OK to apply the style to your map.

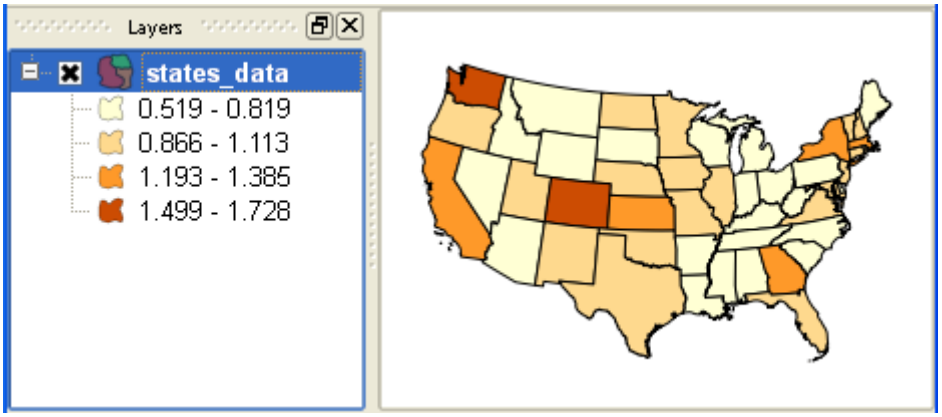


6. *Compare equal intervals to quantiles.* In the symbology tab, we used the default classification scheme called Equal Intervals. This took our four classes of data and divided it so that each class has an equal range of values; with a min value of .519 and a max value of 1.728 our data has a range of 1.209 - divide by four and each class covers a range of .302, sorted from lowest to highest. However, we could use an alternate classification method called Quantiles - this will divide our data into classes that have an equal number of data points. Since we have 51 data points (50 states plus DC), we would have about 13 states in each class sorted from low to high. Double click on the states_data layer to go back to the symbology tab under the properties menu. Change the classification mode to Quantiles and hit Classify. Unfortunately the color scheme reverts back to the original; for each class reset the color to the correct custom color for that class. Hit OK and take a look at the reclassified map. Relative to the equal intervals map, the quantiles show us a greater range of colors as each class has the same number of features.



7. *Classify data using natural breaks.* The natural breaks method classifies data based on the location of gaps or breaks in the data range, which is less arbitrary than equal intervals or quantiles. Some GIS software packages have natural breaks as a classification option; QGIS did not have this option for quite awhile but it is available in version 1.6 under the New symbology tab. Using the old tab or previous versions of QGIS, the simplest way to create natural breaks is to create a scatterplot of your data in a spreadsheet or statistical package and create groups based on visual inspection of the plot (alternatively some stats packages include an algorithim for calculating breaks). Open the symbology tab for the states_data layer. In the classification area, double-click each class and manually change the range based on the plot below. Your four classes should be: 0.519 to 0.819, 0.866 to 1.113, 1.193 to 1.385, and 1.499 to 1.728. Click OK when you're finished and view your map.





8. *Save your project*. At this point  save your project. For our map we'll stick with the natural breaks method, but read the commentary below for an explanation of each method and it's advantages and disadvantages.

Commentary

Data Classification and Color Schemes

The purpose of a thematic map is to communicate a message about the data. If a map uses too few classes, then the data is too generalized and meaningful patterns can be hidden. If a map uses too many classes, then a pattern becomes difficult to detect because there is too much detail. It is difficult for the human eye to

distinguish between too many colors or variations of color. Generally speaking, it is a good idea to use 3 to 6 classes, and ideally 4 or 5. When choosing the number of classes you should consider the number of data points, the range of the data, the purpose of the map, and the color choice based on the output. While a certain number and range of colors may look good on a color printed map, they may appear washed out if the map is shown on a projector or blurred together if photocopied in black and white. You should design with the final output in mind.

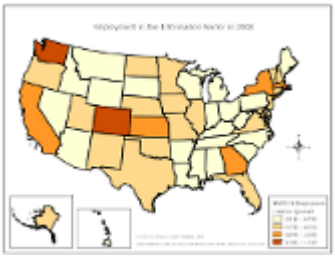
After ranking the data from lowest to highest values, there are a number of classification methods:

- Quantiles - each class has the same number of data points. Always produces distinct map patterns, but can often create categories that have an inconsistent range of values.
- Equal Interval - each class has the same range of data values. Easily understood by map readers, but does not account for data distribution and can result in categories with few or even no values.
- Natural Breaks - classes are created based on the location of gaps in the data. Since the data is divided based on its distribution it is good for distinguishing patterns, but creating the classes is more labor intensive than other methods.
- Unique / Manual - classes created based on some external criteria. Should only be used when justified, otherwise the classification is completely arbitrary.

It's often necessary to make some common sense adjustments to any classification scheme, such as creating unique classes for values of zero or missing values, and adjusting classes so they don't contain a mix of negative and positive values.

The natural breaks method tends to be preferred by geographers for classifying data for maps. Take a look at the maps and data for this project below to compare how the different classification methods group the data; lines under values denote a break between one group and the next. In this particular case the equal intervals and natural breaks method yield almost the same result; this is coincidental. The data we're examining is rather evenly spaced around the mean, and the split based on equal values and the location of gaps in the data is almost identical. In other cases these classification methods will yield quite different results.



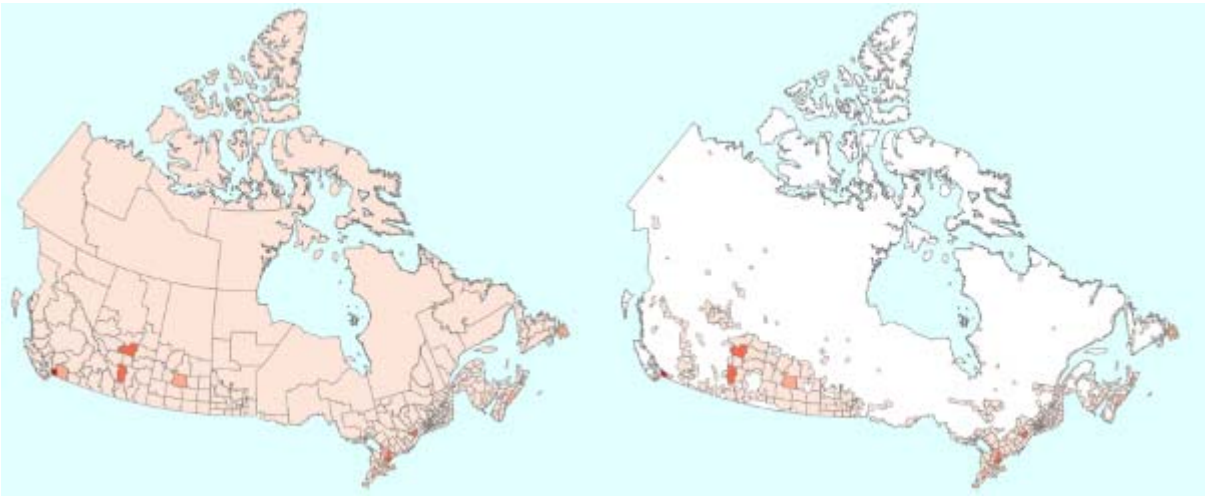


State	Equal Intervals	Quantiles	Natural Breaks
NV	.519	.519	.519
MS	.573	.573	.573
IN	.608	.608	.608
AL	.651	.651	.651
MI	.668	.668	.668
WY	.678	.678	.678
LA	.700	.700	.700
SC	.702	.702	.702
OH	.726	.726	.726
AR	.727	.727	.727
AZ	.728	.728	.728
DE	.726	.726	.726
WV	.738	.738	.738
KY	.756	.756	.756
HI	.764	.764	.764
WI	.795	.795	.795
SD	.806	.806	.806
MT	.809	.809	.809
PA	.811	.811	.811
NC	.812	.812	.812
ME	.813	.813	.813
TN	.817	.817	.817
ID	.819	.819	.819
IL	.866	.866	.866
VT	.872	.872	.872
NH	.874	.874	.874
OK	.884	.884	.884
FL	.900	.900	.900
MD	.912	.912	.912
NE	.927	.927	.927
TX	.943	.943	.943
MN	.946	.946	.946
NM	.948	.948	.948
OR	.950	.950	.950
ND	.983	.983	.983
CT	.992	.992	.992
RI	.995	.995	.995
IA	1.008	1.008	1.008
NJ	1.045	1.045	1.045
MO	1.066	1.066	1.066
UT	1.094	1.094	1.094
VA	1.109	1.109	1.109
AK	1.113	1.113	1.113
MA	1.193	1.193	1.193
GA	1.236	1.236	1.236
KS	1.317	1.317	1.317
CA	1.359	1.359	1.359
NY	1.385	1.385	1.385
CO	1.499	1.499	1.499
WA	1.635	1.635	1.635
DC	1.728	1.728	1.728

Color schemes for displaying quantitative values on choropleth (shaded area) maps should show a logical progression of color values. The progression from light to dark helps convey the change in data values from low to high, and most map readers can infer this without even looking at the map legend. Creating a mixed, fruit salad of colors will defeat this natural inference and will confuse the map reader - so don't do it. When comparing qualitative values (categorical data instead of ranges of values), a map should use colors that reflect

those values. For example, it makes sense to use reds and blues to show which party a state voted for, as these colors have become associated with the US political process. Without even looking at a legend or description, the average American will instantly understand what this map is about. Depicting the same data with greens and yellows doesn't make much sense, and results in confusion.

While we're not considering it for this exercise, the unit of geography used to map phenomena can profoundly affect the interpretation of a distribution or pattern and the ultimate message that your map sends. Mapping populations of US states or Canadian provinces is fine if you are interested in seeing which states / provinces have the most people. But these maps tell you very little about how the population is distributed across these countries, since there is considerable variation in the concentration of people in each state / province. Using a smaller unit of geography can give you a better idea of the distribution of the population. We can see in the first map below that Canada's population is highly concentrated in certain metropolitan areas. However, even the census divisions in the map are not evenly populated. Given that Canada has large unpopulated areas, Statistics Canada has created a layer called an ecumene to show where concentrated areas of population are - this is what you see in the 2nd map. (Source: Geography Division, Statistics Canada, Population Ecumene Census Division Cartographic Boundary File, 2006 Census 92-159-XWE/XWF):



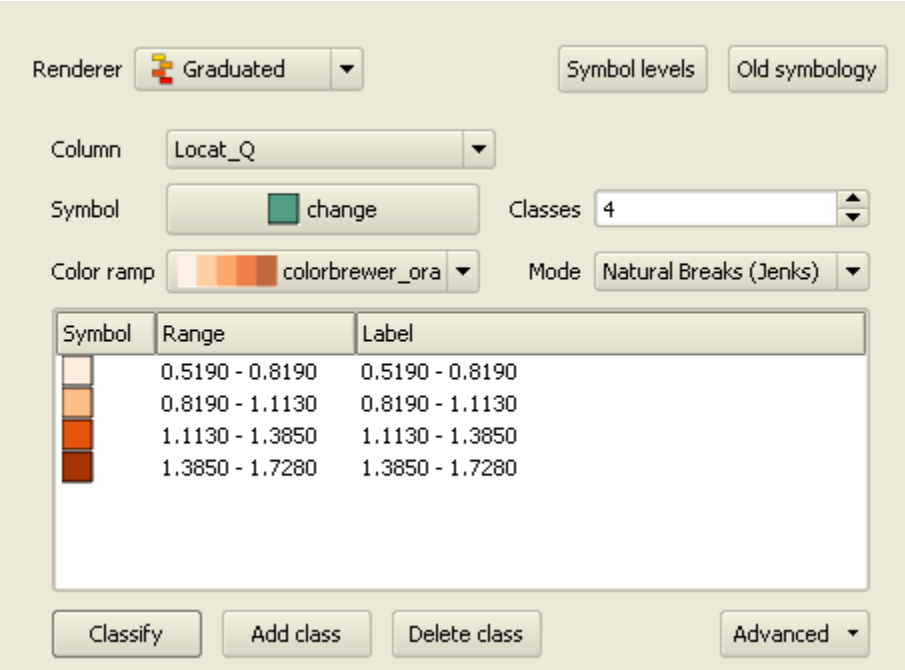
Oftentimes you'll be limited to using certain geographic units based on the availability of the data. For example, it's relatively easy to get current US Census data at the county level, but is rather difficult to get it for zip codes, making it necessary to compromise.

New Symbology Tab

As of version 1.6, the new symbology tab streamlines much of the symbolization process. Access it by pressing the New Symbology button in the current symobology tab. Instead of adding colors manually from Colorbrewer, each of the Colorbrewer schemes are available within QGIS; under the color ramp dropdown you add a new color scheme and select one from the list; you should do this rather than use the default QGIS color ramps, because those ramps are actually backwards (color values go from dark to light for low to high values,

which is the opposite of standard convention). In addition, if you choose to map a different variable or change the classification method you can easily reset all the color values at once rather than having to do it by individual classes.

The new version also includes Natural Breaks and Standard Deviation as classification methods, which means you no longer have to create these manually (although it's still a good idea to examine your data to understand it's distribution).

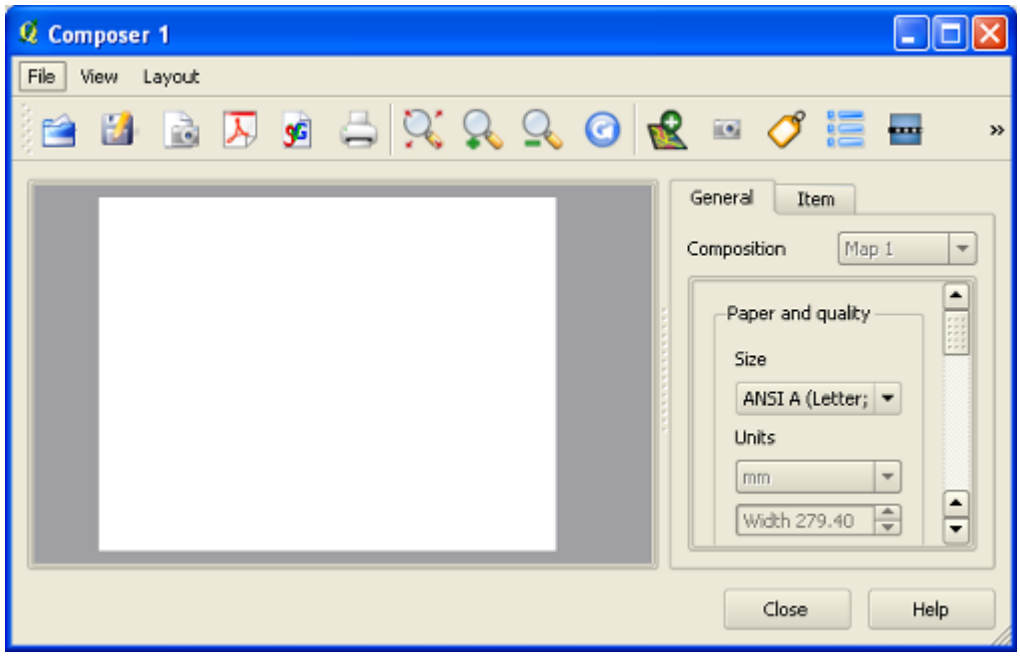


Section V: Designing Maps

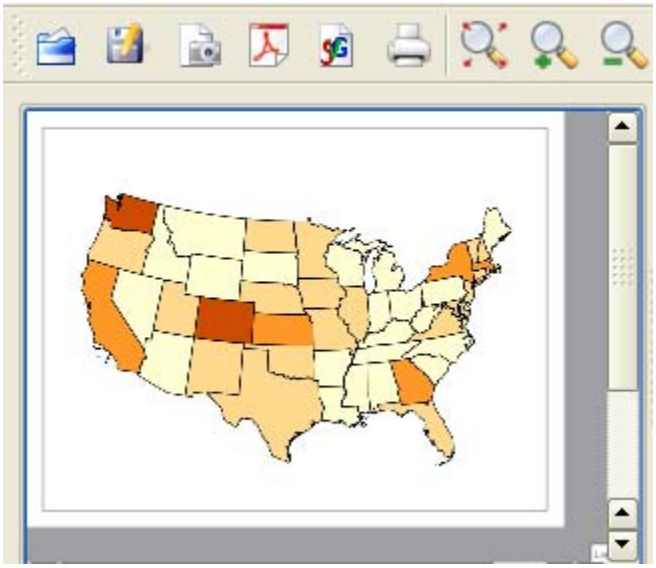
In this section you'll learn how to create a finished map that includes typical map elements: legend, title, scale, and source information.

Steps

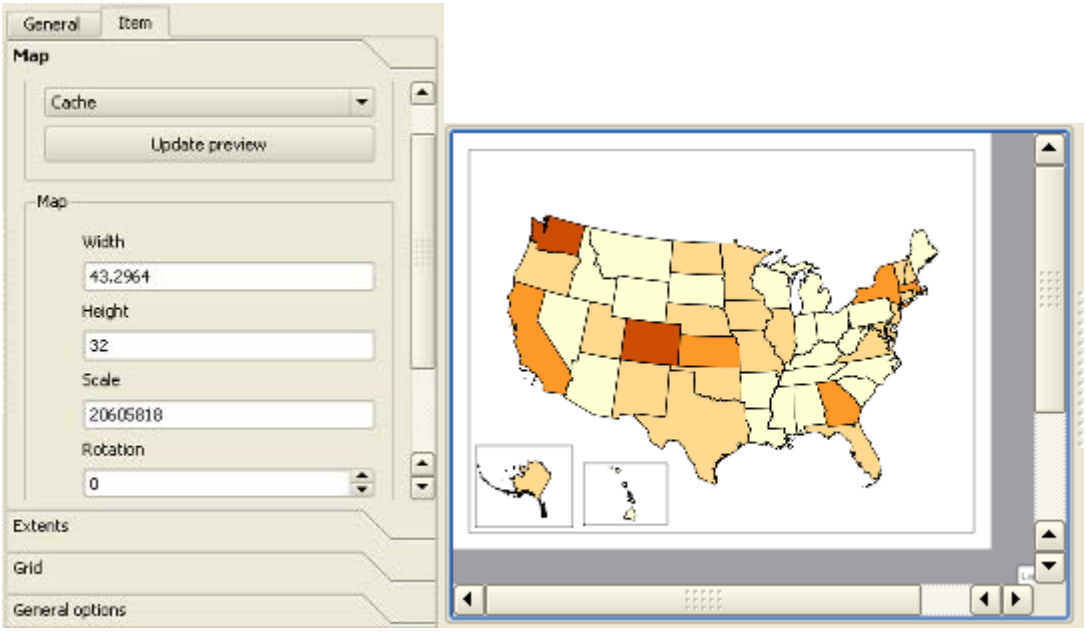
1. *Set the environment for the print layout.* Hit the  print new button to enter the print layout screen. On the General tab in the Paper and quality menu on the right-hand side change the paper size from A4 to ANSI A (letter 8 1/2 by 11). The general tab provides you with options for the map canvas as a whole. Once you add individual items (a map, label, legend, etc) an item tab will appear, and if you have the item selected in the canvas you'll be able to edit its attributes in the item tab.



2. *Add your map and configure zoom.* Hit the add map button in the toolbar. Then draw a box on the map canvas, leaving an even amount of space on each side so there is a gap between the map and the edge of the canvas. If you don't get it right on the first try, you can always hover over an edge of the map, hold down the left mouse, and drag the edge to change the size. Or, to shift the entire map on the page, use the Select Move button. This button moves the entire map box. To shift the geography *inside* the map box, use the adjacent Move Item button. Move the map around so that the lower 48 states are roughly centered in the box. With the move item button selected, you can also change the zoom of the map by using the mouse wheel, or by clicking on the item tab on the left and experimenting with the scale in the map box. The regular zoom buttons on the toolbar will NOT effect the zoom of the geography; these zoom buttons just zoom you closer and further from the map canvas, similar to taking a piece of paper and holding it closer or further from your face. Experiment with them and see. When you're finished, with the map selected go to the Item tab, and under general option increase the outline width of your map from .3 to .5.

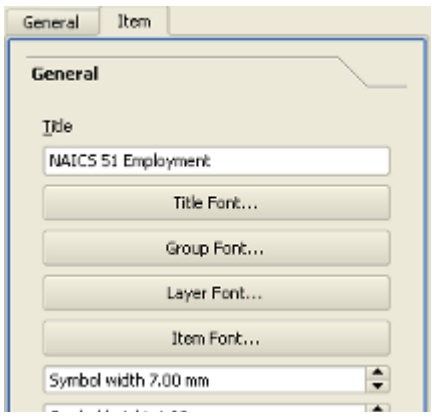


3. *Add additional maps for Alaska and Hawaii.* Given the vast distances between the lower 48 states, Alaska, and Hawaii, it doesn't make sense to include them in the same map window at the same scale; look at most maps of the US and Alaska and Hawaii appear in separate maps or boxes so that optimal scale can be achieved for all three areas; we'll do the same with our map. Hit the  add map button and draw a smaller box in the lower left hand corner. Use the  Move Item button to shift the focus of the map to Alaska, and with this button selected use the map wheel to change the zoom. If you have trouble getting the zoom "right", open the map menu on the item tab on the left, watch how the scale changes as you zoom in and out with the mouse wheel, type in an estimated scale that's somewhere in-between. Once you're finished, repeat the same step for Hawaii and zoom in to focus on the main eight islands. Unfortunately, both AK and HI are going to look a little skewed because they lie at the edge of our continental North American map projection.

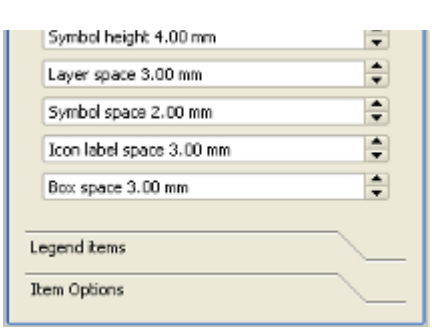


4. *Add a legend.* Hit the  Add Vector Legend button and click on the lower right-hand corner of the map. With the legend selected, hit the Item tab and the Legend Items menu. Select states_data in the list, hit the edit button, and change the name to Location Quotient. You could also edit each data range to change the label (to round numbers or add commas to whole values), but our labels are fine so we'll leave them alone. Stay on the item tab, but flip to the Items Option menu. Then switch to the General menu on the Item tab and change the generic Legend title to NAICS 51 Employment. Hit the Title Font button and change the font to 12. Change the Icon label space and the Box space from 2 mm to 3 mm.

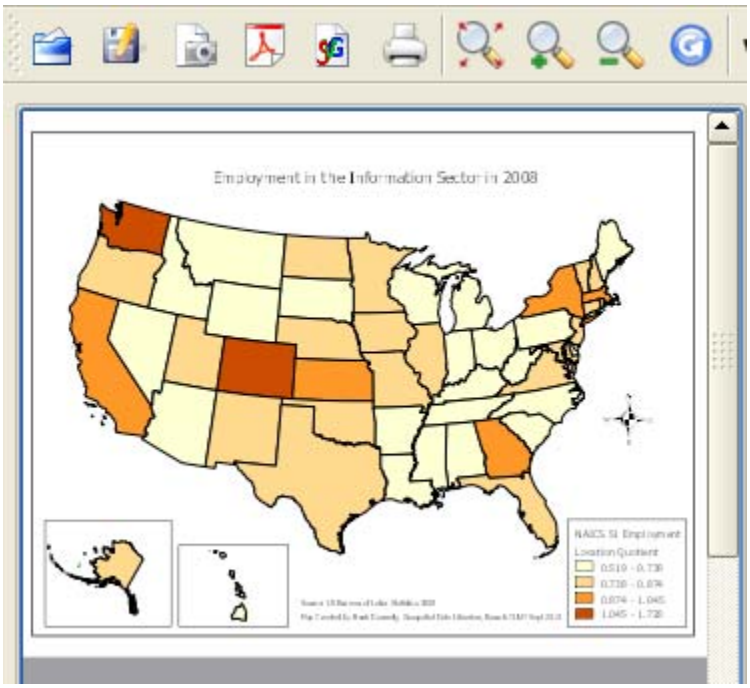
5. *Add a title.* Hit the  Add label button. Click on the top of the map, and a generic label is added. In the label Item tab, change the default Quantum GIS label to Employment in the Information Sector in 2008. Change the font to 18 using the font button. On the Item tab open the General options menu and uncheck the option that says Show Frame. This will turn off the label outline. Click on the label in the map, and using the  select move button, move the label to the top center of the map, and expand the size of the label box so the



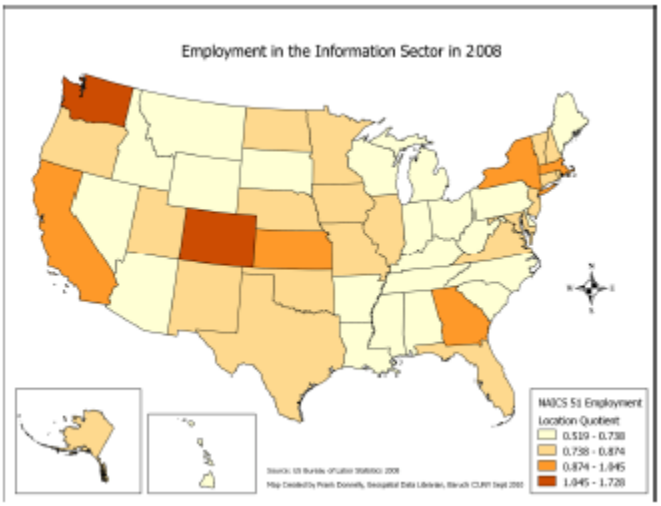
title appears on one line.



6. *Add a label with source information.* Hit the add label button. Click on the bottom of the map to add the generic label. In the label Item tab, change the label to read: Source: US Bureau of Labor Statistics 2008. Change the font to size 8. On the item tab open the General options menu and uncheck the option that says Show Frame. Click on the label in the map, and using the select move button, move the label to the bottom center of the map, and expand the size of the label box so the text appears on one line.
7. *Add a label with author information.* Repeat the same step above to add a label with your information - Map created by [insert your name / organization] [insert date]. Move this label underneath the source label.
8. *Add a north arrow.* Hit the add label button. Click in the upper right hand corner of the map. Change the font to ESRI north and the font size to 72 and hit OK. Erase the default text and type a period "." This will give you a standard compass (you can experiment by hitting other keys to see different arrows). In the item tab, go to the general options and turn the frame for the arrow off.
9. *Did adding the north arrow not work?* The previous step will only work if you have the ESRI fonts stored on your computer. If you don't have those fonts (and you don't feel like downloading them), the other approach would be to add an image of a north arrow. Hit the add image button. Browse to you data folder for part 4 and select arrow.gif. This adds a picture of an arrow to your map. You can try looking for arrows by doing image searches for north arrow or compass using a search engine (Google, Yahoo, Bing, etc).
10. *Balance your map elements.* At this point you should have all of your map elements in place. You may need to resize and shift elements around in order for the map to appear balanced. If you want to insure that boxes are lined up properly, you can hit the Select Move button and click on individual features while holding down the shift key to select multiple items. You can use the various align buttons to arrange elements in a certain way, and you can use the group button to bind several features together so you can move them in unison.



11. *Close the composer and save.* Oddly, there is no save button within the composer (the one on the toolbar is for saving a template of your map, and not the map itself). Close the composer window, and back out at your map view  save your project. This will save the map you just created. It's IMPORTANT that you save your map prior to printing or exporting it - this insures that if the export or print goes wrong or crashes, you won't lose your map. Once you save, hit the  Print button, select the first composer from the list and hit show, and you'll be back to your finished map. If your map looks grainy or out of focus, don't worry - it's really ok. To assuage any worries, you can hit the  Refresh button.
12. *Print to PDF.* PDFs are good for maps that stand alone as their own document. Unfortunately, the  export to PDF function is buggy and often doesn't produce the desired result (see example below). Your other option is to hit the  print button and use a PDF print driver. If you have Adobe installed on your machine Adobe PDF should appear in your list of printers. If not, you can download one of many free print drivers (like Primo PDF) or if you're using a non-Windows machine you can use their preinstalled PDF driver (many linux distros use Evince). There's still a trick to get this to work properly - before your print, use the  Move Item button to shift the geography of the 48 states ever so slightly. Then, with this map still selected, print it to PDF. Save the file in your part 4 data folder as infoemp_2008.pdf.
13. *Export as PNG.* You can also save your map as an image file like a jpg or png. Normally we would want to design the map to be the size of the desired image, and we'd want to adjust the dpi quality in the General tab to reduce it's size. But for now let's just give it a try. Hit the  Save As Image button. Browse to your data folder for part 4 and save the map there as infoemp_2008.png - MAKE SURE you type out the extension .png after the filename in the filename box - otherwise QGIS could freeze completely and you'll be forced to bail out. After you hit save, your screen will flash as it exports - just wait for it redraw and the export will be finished.



14. *Take a look at your maps.* Minimize QGIS and use your file browser to go to your part 4 data folder. Double click on the PDF file to open it in Adobe or your PDF viewing software. If it appears incorrectly (i.e. some of the geometry from the states are missing), go back to the previous step and try the "shifting geography" trick. Double-click on your png file to open it in your default image viewing program (or open it with your web browser). You shouldn't see any problems here. Congratulations on creating a finished map!

Commentary

QGIS Map Composer: Some Details

In some GIS software packages the current view in the map window and the print layout are dynamically linked, and a change in one (such as adjusting the zoom) affects the other. This isn't the case with QGIS; the two are separate. If you do change something in the map view, such as reclassifying the data, you can update the map composer under the item tab for the map by hitting the Update Preview button. Changes in focus or zoom between the view and the composer are not connected at all, which relieves a lot of potential headaches.

The print composer allows you to customize minute details of the canvas, map, and legend, more so than other open source packages. The composer also gives you the ability to  draw shapes or  add portions of an attribute table directly to a map. One of the most recent developments is the ability to store more than one map in a single project. From the map view, you can use the  Print New button to create new, individual

maps, and the  Print Composer button to manage your maps and choose a particular one to show or edit.

The composer also has it's quirks; the lack of a built in north arrow function in the print composer is one that we've seen, and the problems associated with exporting or printing maps to a PDF is another. Some GIS packages allow you to change the map projection for maps within the print composer; this comes in handy for displaying inset maps (like Alaska and Hawaii) correctly, but isn't something that QGIS currently supports. A big weakness is the scalebar feature. While QGIS does allow you to  insert a scalebar, the tool is difficult to use. The scalebar automatically takes the units of measurement used in the map, and there isn't a way to convert units on the fly. So you'll have a scalebar in meters, feet, or decimal degrees instead of kilometers or miles, which is of little practical use. This isn't a large issue if you're creating thematic maps, but if you're designing reference maps not having a scalebar is a problem.

General Map Design

When creating maps you need to design with the end use, format, and audience in mind. If you're designing a map that you're going to embed as an image in a document or web page, you should change the size of the canvas and design the map to the specifications for the document. Creating a full size 8 1/2 by 11 map and scaling or cropping the final image is a bad idea; you'll introduce distortion into the map and text will become illegible. You also need to think about page orientation; it's appropriate to map the United States using a landscape page layout, but if you were mapping an area that was taller rather than wider (South America) you'd want to flip the page to portrait.

Individual map elements (maps, title, arrow, legend, source text) should be balanced on the page to achieve some harmony; avoid lumping too many elements together or having large areas of white space. The title and legend should concisely and accurately describe what the map is about and what you are mapping. The amount of detail you provide and the terminology you use should vary with your audience; for example if we were going to circulate this map to the general public we may want to include a brief explanation of what a location quotient is. You should always include the source of your data in the map. The fonts and north arrows should also be tailored to the map content; a title in calligraphy font and an ornate compass rose may look good if you're recreating one of Christopher Columbus' charts, but it would look rather stupid on our US location quotient map. This may seem like an obvious thing to point out, but the internet is rife with bad maps where people have done just that.

Maps are a method of communication, designed to send a message. Like a book or article that is poorly written, maps that are poorly designed will fail because they do not effectively communicate their message to their audience. Some reasons why maps can flop:

- Poor layout - map elements arranged (map, legend, title, text) in an uneven or sloppy way
- Poor use of symbols - circles too big or small, not enough dots per person, etc
- Improper data classification - too many or few classes that obscure patterns, illogical scheme for dividing data
- Violation of basic cartographic convention - improper conventions for labels and color
- Poor figure-ground relationship - inability to clearly distinguish land from water or foreground from background
- Information overload - too much information (several variables or map elements) or noise (unnecessary information)
- "Chartjunk" - concept defined by the graphic designer Edward Tufte, refers to kitschy or gimmicky elements that add nothing to the message of a map or graphic
- Factual errors - mistakes with labels, data, or geography
- Violates expectation of the user - simplification or generalization is too much for the user to accept
- Offends culture of the user - the message or how the message is communicated (text, colors) violates taboos that a

user or group cannot accept

Output Formats

PDFs are a good format for creating stand-alone documents. PDFs are also a vector-based file, meaning that the geometry of every shape (point, lines, and polygons) is stored as a series of coordinates. If you're working with vector features to begin with, the output in the PDF should be fairly smooth, and if you zoom in to the document you should see all of the detail stored in the original file. The problem with PDFs is they are stand-alone; SVG files are emerging as a vector format that can be embedded in documents but it is still not widely supported (the SVG export in QGIS is also a work in progress).

Image formats are raster-based, meaning that the image is composed of individual pixels or grid cells. Rasters are designed for a specific scale; zoom in too close and the image quality deteriorates as each individual cell becomes more distinct. Raster's can stand alone or can be embedded in documents. PNG files are an open format, compressed raster. They're a good alternative to jpgs; they have better image quality and are widely supported. Tif files are a lossless, uncompressed format - use these only if you need to preserve the image at its highest quality (these files get pretty big). When exporting to a raster, be sure to adjust the dpi (dots per inch) setting, which will adjust the resolution of the image (and affect it's size and quality).

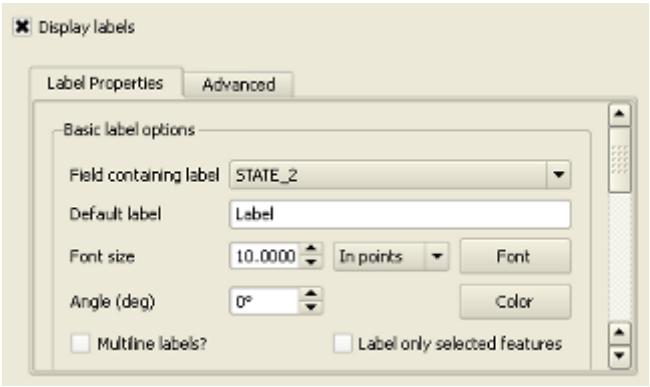
When printing hard copy maps, what you see on the screen is not exactly what you'll get on paper, so be prepared to print test copies and go back and revise. Because there are different screen resolutions and different printers (in terms of print method and quality) colors and outlines will vary. The current ink levels in the printer will also have an impact on how bright or dull the final output is.

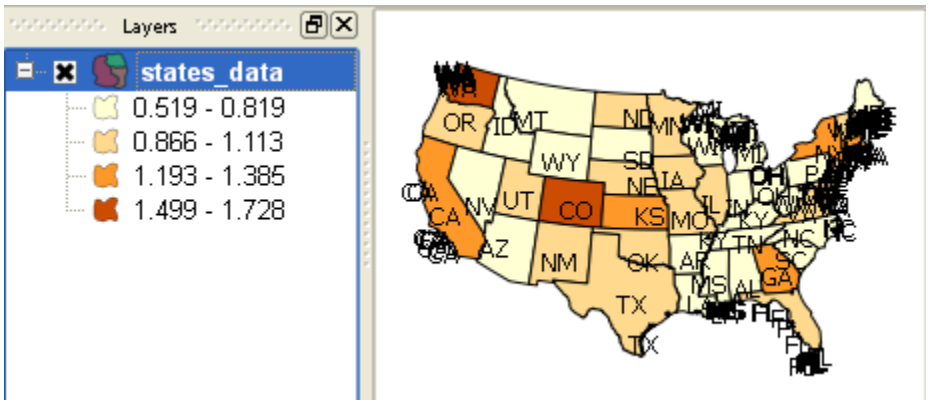
Section VI: Adding Labels

In this section we'll go back and add some labels to our map. The labeling features in QGIS are still a work in progress, but there are some simple work-arounds that you can use to label your features.

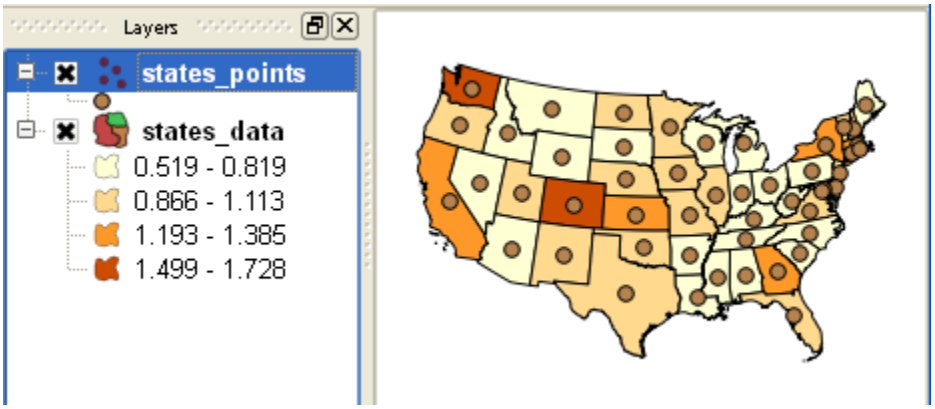
Steps

1. *Turn labels on.* Go back to your QGIS map view. Select states_data in the ML, right click and under the properties menu go to the labels tab. Check the box thats says Display Labels and in the drop down box under field contains labels select STATE_2 as the label field (these are the two letter postal codes for each state - you could confirm this by looking at the attribute table). Change the font to size 8 and click OK. The result is less than ideal; QGIS is assigning a label to every individual polygon, which is not want we want. Go back to the labels tab and turn the labels off.





2. *Create a point layer.* To circumvent this problem, we'll create a point layer out of our states, and use that point layer to place our labels. On the menu go to Vector > Geometry Tools > Polygon Centroids. Select states_data as the input layer, browse to your part 4 data folder and save the new file as states_points. Click yes to add the new layer, then close the menu.



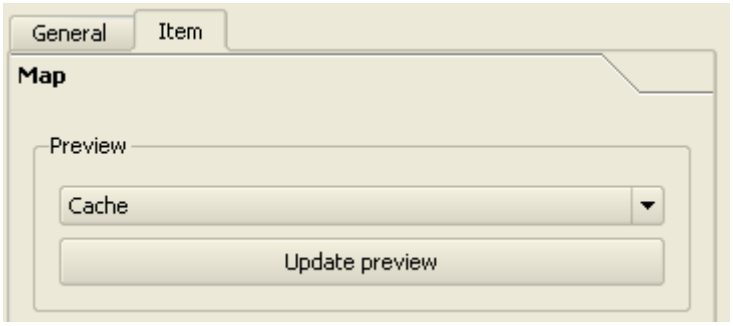
3. *Label the points and remove the symbols.* Open the properties tab for the states_points. Under labels, check the box to turn the labels on, select STATES_2 as the label field, change the font size to 8, and click Apply. Then switch to the symbology tab, scroll down the menu, and under Fill options change the dropdown to none, and under outline options change the outline to none; we want to see just the labels and not the points. Click OK. At this point your labels should be looking better.

4. *Edit your point / label layer.* Some of the labels appear off center (based on how QGIS calculated the center of the polygon) and others are over written (small states in crowded areas). Let's fix these. Select the states_points layer in the ML and hit the  edit button to launch the edit mode. Then select the  move feature button. Hover the crosshairs over the label for Michigan (MI), left click and hold the mouse, drag the label to the center of the state, release the mouse, and the label should be centered. Then re-arrange the labels for Washington DC and Maryland so they do not overlap - try offsetting the DC label so that it is to the left of the district.

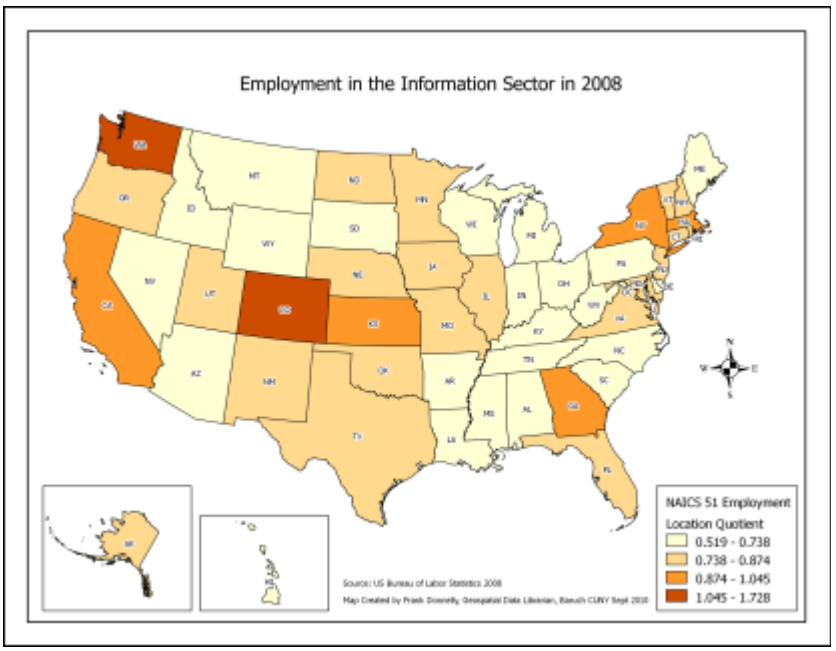
5. *Finish editing and save your edits.* Check out the rest of your map and move additional labels that need to be re-centered (suggestions: look at LA, FL, DE, NJ, RI, and MA). If a label doesn't fit within a state legibly, try offsetting it just outside the state. When you're finished, hit the  edit button and save the changes.
6. *Add buffers to your labels.* If you want your labels to "pop" a bit more, go back to the labels tab under the properties menu for states_points. Scroll through the menu and check the box that says Buffer Labels, and hit OK. If you like the

effect, keep it. Otherwise you can turn them off. At this point Save your project.

7. *Update your map composer.* Hit the print button and show Composer 1. Select the map and under the Item tab hit Update Preview. You should see all your map labels - don't worry if they appear overlapped; they should turn out fine in the export. Repeat this step for the Alaska and Hawaii map box. Then select the legend, select states_points in the list in the Item tab under legend Items, and hit the X button (beside the edit button) to remove it from the legend (we don't need this to appear).



8. *Save and export.* Close the map composer and back in the map view hit the save button. Then go back to the map composer and export your map (remember - we exit, save, and return just in case the export crashes). Print your map out as PDF or save it as an image. Save it in your part 4 data folder as infoemp_2008_labels.png (or .pdf). Minimize QGIS, go to your part 4 data folder, and take a look at your final map.



Commentary

Labeling in QGIS

Features can be displayed and differentiated from each other using text. For example, the standard cartographic convention for labeling oceans is to use an italic font and, when possible, a dark blue color. The size of a label indicates the hierarchy of the feature - oceans have larger fonts than seas, which have larger fonts then rivers, larger than streams, etc. Land features are labeled in black, or anything that isn't blue, and are never

written in italics. Larger features, land or water, may be written in all capital letters, while smaller features are in lower case.

ATLANTIC OCEAN *GULF OF MEXICO* *Lake Ontario* *Hudson River*

UNITED STATES NEW JERSEY Philadelphia Trenton

Currently the automatic labeling placement in QGIS leaves something to be desired and we've demonstrated how you can get around this issue by creating a point layer that serves as a dedicated labeling layer. Other options at your disposal:

-  The new, experimental labeling tool is available via the map view. It does a a better job than the default label options, but is still a work in progress.
-  The text annotation tool allows you to add call out boxes directly in the map view. This is practical if you only need to place a few labels.
-  You can also use the add label feature within the map composer. This can be a little cumbersome since you cannot copy and paste labels, but must create each one from scratch; ok if you only need to add a few labels.
- You can add columns to your attribute table that allow you to specify label details for each feature. This allows you to use the advanced menu under the labels tab in the properties menu to designate your annotation fields. For placement you can define angles or provide coordinates (like latitude and longitude) to specify where labels can be placed. You can also have columns to indicate font type, size, color, etc. This is worth the effort for longer term projects.

Thematic Maps and Symbols

In this tutorial we worked through an example for creating a shaded area or choropleth map. However, there are a number of other techniques that you can use to create a thematic map. QGIS also supports graduated symbols for point and line layers, where the relative size of the symbol (a circle, square, line, or image) represents a value (if you look at the symbology tab for a point layer, you can change the legend type to graduated symbols). If you have a polygon layer that you'd rather map as graduated circles (instead of shaded areas) you have to convert it to a point layer first.

Symbols are used to show qualitative data (name or feature type) or quantitative data (proportions or numbers) and are often divided into four types:

- Nominal - qualitative measurements like the name or type of feature, shown using unique symbols.
- Ordinal - quantitative measurements with a general order of size, like small, medium, or large, shown using symbols of different sizes or colors.
- Interval - quantitative measurements with a specific beginning point and range of specific values (distance, temperature, elevation), shown using a variety of symbols (isolines, shaded areas, graduated symbols).
- Ratio - a type of interval measurement that shows the relationship between the area and some phenomena (time to cover a distance, population density).

Symbols are often designed to mimic the features they represent, i.e. airplanes for airports, little buildings with flags to represent schools, etc (these are all examples of nominal symbols). In some cases, features may be represented with geometric shapes (circles, squares, triangles) that can be easily distinguished on small scale

maps. Some features may be represented using a standard convention for classifying them, i.e. mining maps may label minerals based on their abbreviation in the periodic table - Sn for tin, Pb for lead, Cu for copper, etc.

A single symbol can be used to identify a feature. Varying the size or color of the symbol can indicate quantity. The width and color of roads on a map is highly standardized to show the type of road and volume - thick blue roads are interstate highways, thick green roads are toll highways, thinner red roads are US highways and thinner black roads are state or local roads (all ordinal symbols).

Considerations and Next Steps

Now that we have mapped this data - what does it mean? How would you interpret this map? Are there any spatial patterns to the data (clustering) or does it appear more or less random? Maps have the ability to answer questions but also raise new ones. In order to understand what's going on, we have to become familiar with the underlying dataset. What kinds of occupations are included in the Information Sector, and how might that explain the distribution across different states?

For more practice, something else to try:

- In addition to shaded areas, we can also create graduated circle maps. Take the label layer that you created, turn the labels off, and symbolize the layer based on the total number of jobs in the information sector by state. Hop into your map layout and create a bi-variate (two variable) map that shows the number of jobs (as circles) and location quotient (as shaded areas).

Part 5 - Going Further

This tutorial has provided you with a basic introduction to GIS concepts and applications using QGIS. This section will cover the next steps you can take on your own.

Section I: Finding Data

Throughout this tutorial you've been provided with data that you've used to work through various exercises. Once you're working on your own projects, you'll need to find or create the data you need. There is a lot of free GIS data available on the web, created by various government agencies, academic and non-profit organizations, and private companies. You can try a search engine or look at an academic map / GIS library website for a list of helpful links (a list of suggestions is included in the following section). To be strategic about your search, it helps to understand who creates and provides the data:

- *Global / international:* Look at supra-national agencies, like the United Nations (in particular, the UN's Environment Programme has a good site) or academic / non-profit organizations who have enhanced and updated public domain data such as the Global Administrative Areas (GADM) site, the DIVA GIS data page, and the Natural Earth project. If you need satellite imagery the best sites to visit are the USGS and NASA.
- *Country level:* In some cases you'll want to visit a few of the international sites, like DIVA GIS and Natural Earth, to get basic country-level datasets like state or provincial boundaries. But in many instances you may want to visit a mapping agency website or data depository for the specific country you're interested in; you'll find more country specific layers and they will be processed in a way that is readily compatible for mapping attribute data from that country. Most countries have one or two agencies that will provide the bulk of the country's GIS data - a statistical agency responsible for the census, or a mapping agency responsible for surveying. In the US you could go directly to the US Census Bureau or the USGS to download data, or you could visit the central data.gov repository. In Canada, you could visit Statistics Canada directly or visit the Geogratis repository. Some countries may provide one central source (Australia), whereas other countries may provide limited or no data, via a website that may or may not be in English.
- *State / Provincial:* You may be able to visit a country level source, like the US Census Bureau to get state, county, or zip code boundaries for the entire state, or you can visit a state level agency to get more specialized datasets for that state. Some states will have state government portals where you can access all data for a state, others may cooperate with a college or university located in that state to provide data via the university's portal. In addition to centralized portals, individual departments or agencies may also provide data directly; road and transportation layers may be provided by a state department of transportation or may be provided through the state's central portal. State agencies are also the most likely source for aerial photography.
- *County / City / Local:* Local governments may have portals where they provide administrative boundaries, transportation data, and real estate or tax parcels, and datasets that would be of local interest (such as neighborhood boundaries that may not be formally defined elsewhere). You can also look at the geography one step above (state level) to see if data is available for the local area.
- *Gazetteers and Geocoding:* if you can't find an existing GIS dataset, you can always try to create one from an online

gazetteer that provides latitude and longitude coordinates for point-based features; the USGS has a US level gazetteer, while the NGA has an international gazetteer. Do you have a list of addresses but no coordinates? Try uploading them to a free geocoding service like the one at USC GIS Research Laboratory, which will translate your addresses into coordinates.

- In some cases you may find university or non-profit sites that provide data within a specialized area of interest. While universities typically provide data for the geographic areas where they reside, there may be special labs or research groups that provide data beyond that area; the CIESN (Center for International Earth Science Information Network) site at Columbia University the NHGIS (National Historic GIS) at the University of Minnesota are two examples.

Regardless of where you download your data, you'll want to examine the metadata for the layers. Metadata can be formally or informally described on the website where you downloaded your files, in narrative documentation that is included with the files you downloaded, or in special XML files that accompany each of your GIS files. There are a few well-defined standards such as the FGDC and ISO 19139 that data creators use to document data, and include elements that explain who created the data, when it was last updated, what the file contains, what the intended purpose of the file is, if it was created for a specific optimal scale, the coordinate system and map projection it was created in, and copyright and use restrictions. You'll want to check the metadata to verify that the data is going to meet your needs and that you can use it for your intended purpose. For example, you wouldn't want to use a generalized boundary file if you're mapping at a large, local scale, and if you are going to use the data for a commercial purpose you need to verify that that's permitted. In any event, you should cite the source of your data in any maps, tables, or reports you create from it.

If you are looking for a particular GIS file and it's provided by several sources, which source should you use? For example, if we wanted census tracts for a particular city, we could download them from the city's GIS page, from a state-based site, from one of ESRI's pages, or from the Census Bureau itself, via the TIGER page or the generalized boundary page. To answer this question, you'll have to examine the download page, and even download the files to view them and their metadata. Here are some things to consider:

- How are the files packaged for download? Do I have to download them one place at a time, or could I get the entire area in one download?
- Who created the files originally? Is it better to go with the original source? Or has a secondary source added some value that makes their files more desirable?

- Can I trust the source? Is there metadata? How did they create the data?
- For vector files, are the layers generalized or not? What scale are they appropriate for?
- For vector files, are the polygons saved as single or multipart layers?
- For vector files, what attributes are available in the attribute table? Are there ID codes that I can readily use to join data? Are there place names that I can readily use as labels?
- For raster files, what is the resolution of the data? Is it appropriate for my intended use?
- What format is the file in? Is it a format I can use, or at least one that I can easily convert?
- Are there any copyright or use restrictions with the data?

Finally, remember that GIS data is often just one piece of the puzzle. It represents the geographic features, but if you need attributes to go with these features (demographic data, weather data, sales data, etc) you'll have to download this data from someplace else (or create it yourself) and process it to make it usable with your GIS data.

Section II: Data Sources

Global

- DIVA GIS data – Country level vector and raster data for every single country in the world – download individual files or geodatabases. Assembled for the BioGeomancer Project at UC Berkeley and part of the DIVA GIS project. For just global administrative boundaries, you could also visit the GADM database page.
- Natural Earth – Generalized raster and vector data for countries, available at three different scales.
- United Nations Environment Program – Geodata Portal. Click on “Advanced Search,” select “Geospatial Data Sets” under the first drop down box, and hit the red “Search” button. This will take you to a list of global or continental GIS files that you can download.
- Center for International Earth Science Information Network – Hosted by Columbia University, it contains links to datasets for the world, various countries, and the US.

Canada

- GeoGratis – Canadian government GIS repository provided by the Earth Sciences Sector of Natural Resources Canada.
- GeoConnections – Portal to a wide range of Canadian geospatial data. Some resources are free, others must be purchased.

United States

- TIGER Line Shapefiles – U.S. Census Bureau – extracts of the bureau’s TIGER Line files for several legal, administrative, and statistical areas in the US, updated annually.
- Cartographic Boundary Files – U.S. Census Bureau – generalized extracts of the bureau’s TIGER Line files for several administrative areas (i.e. states, counties, zip codes) and census (i.e. tracts, block groups, metros) areas in the US.
- National Historical Geographic Information System – the NHGIS is a project at the University of Minnesota that compiles and provides historical census boundaries and data for the United States from 1790 to 2000. New users must register, but there is no cost and downloads are free.

- Data.gov's Geodata Catalog – large depository of GIS data from several federal agencies.
- USGS Seamless Data Distribution System – this federal agency provides imagery, digital topographic maps (DRGs), elevation data, and some boundary files.

State of New York

- CUGIR – Cornell University's Geospatial Information Repository. They also compile data at the state, county, and local levels for NY State and they coordinate their activities with NYS GIS.
- NYS GIS – Digital Orthoimagery Direct – the NYS GIS page for imagery (orthophotos), tiles can be searched by county and year. Imagery for the five boroughs for the most current series is only available by direct, special request. Imagery from the older series is available for all areas.

New York City

- NYC Data Mine – this site is a repository of geospatial and attribute data from several city agencies.
- BYTES of the BIG APPLE – the NYC Department of City Planning's page has administrative and political boundaries, streets, transportation networks, shorelines, and tax parcels.
- DoITT – Services: GIS – the NYC Department of Information Technologies and Telecommunications has transportation networks, survey points, water bodies, building footprints, and open spaces.

Section III: Additional Concepts and Applications

In this tutorial you've learned what GIS is, what it looks like, and generally how it works. You've learned how to work with vector-based GIS data to do some basic geoprocessing and analysis, and you've learned the basics of thematic mapping and map design. Here are some things that we didn't cover that you may wish to explore next, using QGIS as your GIS application.

- *Working with rasters*. The addition of the GDAL plugin allows you to do more interesting things with rasters that were previously not available in QGIS.
- *Creating and editing vector layers*. QGIS has an entire suite of tools that allow you to edit files point by point, line by line, feature by feature, and to create files from scratch.
- *Georeferencing*. The georeferencing plugin gives you the ability to take non-GIS raster files (a map or chart in a jpg or basic image file that lacks coordinates) and transform it into a GIS layer.
- *Geodatabases*. Instead of storing all of your features in individual shapefiles and your attribute data in several DBFs, store everything in a single database file. Use the database software to organize your data to run spatial and non-spatial queries. QGIS can directly connect to the desktop Spatialite database or the server-based PostGIS database.
- *WMS and WFS*. Tap into server and web-based datasets without downloading anything. Data provided in a WMS (web mapping service) format can be displayed as a raster in QGIS, while WFS (web feature service) layers can be viewed as vectors via a plugin.
- *Learn command line tools*. Need to export data from one format to another? Or reproject files? Or rename them? Do you have large batches of files to change or transform? The GDAL / OGR tools, many of which are embedded in QGIS, are also available via the command line or shell and can make your life a little easier.
- *Need more analytical capabilities?* The are a number of other vector analysis tools under the ftools menu, but you

can also try the QGIS GRASS plugin, and learn how to use the powerful GRASS GIS software. The learning curve is a bit steep, but with the GRASS tools you'll have more than enough features to match the major proprietary software.

The QGIS website and the OSGeo foundation have links to additional manuals and tutorials for learning QGIS and GRASS. In print, Sherman's *Desktop GIS: Mapping the Planet With Open Source Tools* is great for delving deeper into QGIS and for providing a crash course in GRASS, PostGIS, and the GDAL OGR command line tools. *Open Source GIS: A GRASS GIS Approach* by Neteler and Mitasova is the definitive source for learning about GRASS. In addition to QGIS and GRASS, there are a number of other open source GIS products bouncing around that are worth a look. gvSIG, an open source desktop GIS package created by local government agencies in Spain, is a notable alternative.

If you think you're going to become deeply involved in GIS, you may want to consider trying the major proprietary packages in the industry such as ESRI's ArcGIS or Pitney Bowes MapInfo. If you're a current Baruch student, faculty, or staff member you can sign up to take free, self-paced, online courses in ArcGIS as part of the ESRI Virtual Campus program. Visit the ESRI VC page under the Tutorials and Courses tab on Baruch GIS Subject Guide (http://guides.newman.baruch.cuny.edu/gis) for information on how to sign up. ArcGIS is available in several computer labs on campus. CUNY students outside of Baruch should contact the site license administrator of ArcGIS on your campus to see who administers the courses to gain access. Once you're familiar with QGIS, the leap to one of the proprietary packages isn't too great because they use a similar interface and operate under the same basic principles. ArcGIS is well documented; there are many books and online tutorials. On the flip side, the software is more resource intensive, is only available for the Windows operating system, and is expensive enough that it's not a viable option for an individual user. You can download and sample a basic, freeware version called ArcExplorer from ESRI's website.